(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 July 2001 (26.07.2001)

PCT

(10) International Publication Number
WO 01/53481 A1

(51) International Patent Classification[7]: C12N 15/10, 15/62, 15/63, 15/64, 15/65, 5/10, C12Q 1/68, G01N 33/50

(21) International Application Number: PCT/US01/01480

(22) International Filing Date: 18 January 2001 (18.01.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/177,250 20 January 2000 (20.01.2000) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/177,250 (CIP)
Filed on 20 January 2000 (20.01.2000)

(71) Applicant *(for all designated States except US)*: **CELL THERAPEUTICS, INC.** [US/US]; Suite 400, 201 Elliott Avenue West, Seattle, WA 98119 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **LOFQUIST, Alan** [US/US]; 9422 N.W. 130 Place, Kirkland, WA 98034 (US). **LEUNG, David** [US/US]; 7625 East Mercer Way, Mercer Island, WA 98040 (US).

(74) Agents: **BENT, Stephen, A.** et al.; Foley & Lardner, 3000 K Street, N.W., Washington, DC 20007-5109 (US).

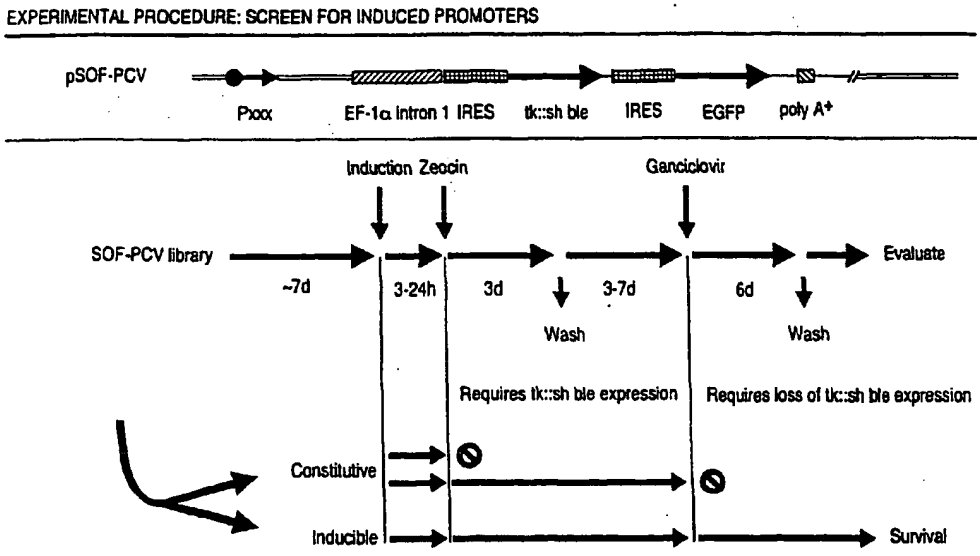
(81) Designated States *(national)*: JP, US.

(84) Designated States *(regional)*: European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHODS AND COMPOSITIONS FOR IDENTIFICATION OF GENOMIC POLYNUCLEOTIDES WHICH ARE TRANSCRIPTIONALLY REGULATED



(57) Abstract: The present invention relates to methods and compositions for the identification of portions of the genome which are modulated by compounds. The invention also relates to methods of comparing the effects of modulating compounds. The invention also relates to a plurality of eukaryotic cells having integrated at a plurality of integration sites the nucleic acids of the present invention.

WO 01/53481 A1

Methods and Compositions for Identification of Genomic Polynucleotides which are Transcriptionally Regulated

Technical Field

The present invention generally relates to methods and compositions for the identification of portions of the genome and compounds for modulating such portions of the genome. The present invention is particularly directed to methods and compositions for the identification of proteins that are directly or indirectly transcriptionally regulated and compounds for regulating such proteins either directly or indirectly.

References

The following references are cited in the application and are hereby incorporated into the application by reference in their entirety.

1. International Patent Application Publication No. WO98/13353

2. Sambrook et at. *Molecular Cloning A Laboratory Manual*, 2d ed. (1989) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

3. Dayhoff, M.O., in *Atlas of Protein Sequence and Structure*, (1972) Volume 5, National Biomedical Research Foundation, pp 101-110, and Supplement 2 to this volume, pp.110.

4. Smith and Waterman (1981) *Adv. Appl. Math.* 2:482,

5. Needleman and Wunsch (1970) *J. Mol. Biol.* 48: 443,

6. Pearson and Lipman (1988) *Proc. Natl. Acad. Sci.* (U.S.A.) 85: 2444,

7. Lupton S.D. et al., (1991) "Dominant positive and negative selection using a hygromycin phosphotransferase-thymidine kinase fusion gene" *Mol. Cell Biol.* 11: 3374-3378;

8. Karreman C. (1998) "A new set of positive/negative selectable markers for mammalian cells" *Gene* 218:57-61

9. Akatsuka, Y. et al., (1994) "Retrovirus-mediated transfer of a hygromycin phosphotransferase-thymidine kinase fusion gene into human CD34++ bone marrow cells" *Int. J Hematol.* 60:251-261;

10. Beck, C. et al., (1995) "The thymidine kinase/ganciclovir-mediated "suicide"effect is variable in different tumor cells" *Hum. Gene Therapy* 6:1525-1530;

11. Veelken H. et al., (1996) "Systematic evaluation of chimeric marker genes on dicstronic transcription units for regulated expression of transgenes in vitro and in vivo" *Hum. Gene Ther.* 7:1827-1836;

12. Harrison R.W. and Miller J.C. (1996) "Functional identification of genes up- and down-regulated by glucocorticoids in AtT-20 pituitary cells using an enhancer trap" *Endocrinology* 137:2758-2765

13. Bonaldo P et al., (1998) "Efficient gene trap screening for novel developmental genes using IRES - beta-geo vector and *in vitro* preselection" *Exp. Cell Res.* 244:125-136;

14. International Patent Application No. W09611211 by Das and Coward published 4/16/96,

15. European Patent Application No. 585983 by Zurr published 3/7/96;

16. International Patent Application No. W096/01324 by Berlioz published 1/18/96;

17. International Patent Application No. W094/24301 by Smith published 10/27/94;

18. Kuspa and Loomis, (1992) *Proc. Natl. Acad. Sci* 89: 8803-8807

19. Derbyshire, K.M., (Nov. 7, 1995) *Gene* 143-144;

20. Zambrowicz B.P. and Friedrich G.A. (1998) "Comprehensive mammalian genetics; history and future prospects of gene trapping in the mouse." *Int J. Dev. Biol.* 42:1025-1036;

21. Cannon J.P. et al., (1999) "Gene trap screening using negative selection:identificaiton of two tandem, differentially expressed loci with potential hematopoietic function". *Dev. Genet.* 25:49-63;

22. *Science* (1994) 264:1415-1421;

23. *Mol. Cell Biol,* (1996) 16:369-375;

24. U. S. Patent No. 5,436,128;

25. Shaw, et al. (1988) *Science* 291:202-205;

26. Simon et al., (1991) *Science* 252:802-808;

27. Flanagan et at., (1991) *Nature* 352:803-807;

28. Watson and Arkinstall. *The G-Protein Linked Receptor Facts Book,* Academic Press, New York (1994);

29. Karin M., (1994) "Signal transduction from the cell surface to the nucleus through the phosphorylation of transcription factors" *Curr. Opin. Cell Biol.* 6:415-424;

30. Levitzki A "Targeting signal transduction for disease therapy"(1996) *Curr Opin Cell Biol.* 8:239-244;

31. Suthanthiran et al., *Am. J. Kidney Disease,* 28:159-1 72 (1996);

32. Lu, *Basic Toxicology, Fundamentals. Target Organs. and Risk Assessment,* Hemisphere Publishing Corp.. Washington (1985);

33. U.S. Patent Nos: 5,196,313 to Culbreth (issued March 23, 1993)

34. U.S. Patent No.5,567,952 to Benet (issued October 22, 1996);

35. Creasey, *Drug Disposition in Humans. The Basis of Clinical Pharmacology*, Oxford University Press, Oxford (1979)

36. Daluge et al., *Antimicro. Agents Chemother.* 41:1082- 1093 (1995);

37. Shaw and Lacy. S. *Bone Joint Surg. (Br)* 55:197-205(1973);

38. McDonough, *Phys. Ther* 62:835-839 (1982);

39. Agrafiotis et al., U.S. patent 5,574,656

40. *Remington's Pharmaceutical Sciences*, 18th ed. Mack Publishing Co. (1990)

41. Fingl et al. *The Pharmacological Basis of Therapeutics*, (1975)

42. Ausubel et al., (1989) "Current Protocols in Molecular Biology" John Wiley & Sons, Baltimore MA;

43. Perbal (1988) "A Practical Guide to Molecular Cloning" John Wiley and Sons, New York.

44. PCR protocols: A Guide to Methods and Applications, Academic Press, San Diego CA. (1990)

45. Weeks et al., *DNA Cell Biol.* 16:281-289 (1997)

46. Leung et al., *Proc. Natl. Acad. Sci, USA* 92:4813-4817 (1995)

47. Uetsuki et al., *J. Biol. Chem.* 264:5791-5798 1989

48. Cachianes et al., *Biotechniques* 15:255-259 1993

49. Whitney et al., (1988) *Nature Biotechnology* 16:1329-1333

50. Benton and Davis (1978) *Science* 196:180

51. Smith and Waterman (1981) *Adv. Appl. Math.* 2:482;

52. Needleman and Wunsch (1970) *J. Mol. Biol.* 48: 443;

53. Pearson and Lipman (1988) *Proc. Natl. Acad. Sci.* (U.S.A.) 85: 2444

Background

The identification and isolation of useful portions of the genome requires extensive expenditure of time and financial resources. Currently, many genome projects use various strategies to reduce cloning and sequencing times. While genome projects rapidly expand the database of genetic material, such projects often lack the ability to integrate the information with the biology of the cell or organism from which the genes were isolated. In some instances, coding regions of newly isolated genes reveal sequence homology to other genes of known function. This type of analysis can, at best, provide clues as to the possible relationships between different genes and proteins. Genomic projects in general, however, suffer from the inability to rapidly and directly isolate and identify specific, yet unknown, genes associated with a particular biological process or processes.

The evaluation of the function of genes identified from genomic sequencing projects requires cloning the discovered gene into an expression system suitable for functional screening. Transferring the discovered gene into a functional screening system requires additional expenditure of time and resources without a guarantee that the correct screening system was chosen. Since the function of the discovered gene is often unknown or only surmised by inference to structurally related genes, the chosen screening system may not have any relationship to the biological function of the gene. For example, a gene may encode a protein that is structurally homologous to the β-adrenergic receptor and have a dissimilar function. Further, if negative results are obtained in the screen, it can not be easily determined whether 1) the gene or gene product is not functioning properly in the screening assay or 2) the gene or gene product is not directly or indirectly involved in the biological process being assayed by the screening system.

International Patent Application Publication No. WO98/13353 provides a method for the identification of genomic polynucleotides using a beta-lactamase expression construct (Whitney et al., Nature Biotechnology Vol 16, 1988 1329-1333). Although such a method and construct is able to select for those proteins whose transcription is either induced or repressed by the addition of a modulator compound, the method taught by International Patent Application No. WO98/13353 requires at least three rounds of repetitive fluorescence cell sorting of the induced or repressed clones. Despite iterative selection using sophisticated flow cytometry instrumentations, the isolated clones still contain a high background of false positives (70% for induced clones, 90% for repressed clones) such that the identification of proteins whose transcription is altered by a modulator compound is difficult.

The present invention is based on a survival gene with both dominant positive and negative selection properties. Survival genes with both dominant

positive and negative selection properties have been described (Lupton S.D. et al., "Dominant positive and negative selection using a hygromycin phosphotransferase-thymidine kinase fusion gene" *Mol. Cell Biol.* 1991: 11: 3374-3378; Karreman C. "A new set of positive/negative selectable markers for mammalian cells" gene 1998;218:57-61).

An example is the hygromycin phosphotranferase gene (hy) fused in-frame with the herpes simplex virus type 1 thymidine kinase gene (TK). The resulting fusion gene (termed HyTK) confers hygromycin B resistance for dominant positive selection and ganciclovir sensitivity for negative selection and provides a means by which these selectable phenotypes may be expressed and regulated as a single genetic entity. It is mainly being used in clinical gene therapy trials as a therapeutic gene or as a safety marker (Akatsuka, Y. et al., "Retrovirus-mediated transfer of a hygromycin phosphotransferase-thymidine kinase fusion gene into human CD34++ bone marrow cells" *Int. J Hematol.* 1994:60:251-261; Beck, C. et al., "The thymidine kinase/ganciclovir-mediated "suicide"effect is variable in different tumor cells" *Hum. Gene Therapy* 1995:6:1525-1530; Veelken H. et al., "Systematic evaluation of chimeric marker genes on dicstronic transcription units for regulated expression of transgenes in vitro and in vivo" *Hum. Gene Ther.* 1996:7:1827-1836).

There has been only one report of using the HyTK fusion gene driven by a minimal promoter for trapping enhancers for the identification of genes that can be up- or down-regulated by glucocorticoids in tT-20 pituitary cells (Harrison R.W. and, Miller J.C. "Functional identification of genes up- and down-regulated by glucocorticoids in AtT-20 pituitary cells using an enhancer trap" *Endocrinology* 1996: 137:2758-2765). However only a small portion of insertion events occurred at different sites.

Consequently, there is a need to provide simpler and more economical methods and compositions for rapidly isolating portions of genomes associated with a known biological process and to screen such portions of genomes for activity without the necessity of transferring the gene of interest into an additional screening system. Furthermore, there is a need to provide methods and compositions for rapidly and consistently isolating portions of genomes in which transcription is either positively or negatively modulated by a known modulator.

This present invention uses a novel set of gene trap vectors to screen and to identify modulator regulated genes. These vectors are based on a survival gene with both dominant positive and negative selection properties placed downstream from a splice-acceptor sequence and an internal ribosome entry site sequence (IRES). The vectors described in the current invention allow trapping of regulated promoters when integration occurs downstream of a promoter of interest as well as within any intron or exon sequences. The present invention further expands the repertoire of survival genes for isolation of modulator regulated sequences.

Brief Description of the Figures

Figure 1A is a representation of the method by which the insertion and expression of the survival genes reports the increase in expression of a pathway within a cell.

Figure 1B is a representation of the method by which the insertion and expression of the survival genes reports the decrease in expression of a pathway within a cell.

Figure 2A is a schematic map of pIRESpuro. Puro^r indicates the puromycin resistance gene, IRES indicates the IRES element (internal ribosome entry binding site), Amp^r indicates the ampicillin resistance gene.

Figure 2B is a schematic map of pIRES2-EGFP. IRES indicates the IRES element (internal ribosome entry binding site), EGFP indicates enhanced green fluorescent protein; Kan^R or Neo^R indicates respectively the kanamycin resistance gene or the neomycin resistance gene or the G418 (geneticin) resistance gene, P_{SV40} in the promoter from SV40, HSV TK is the Herpes simplex virus thymidine kinase gene.

Figure 3A is a schematic map of pFrog-CMV. IRES indicates the IRES element (internal ribosome entry binding site), EGFP indicates enhanced green fluorescent protein; Kan^R or Neo^R indicates the kanamycin resistance gene or the neomycin resistance gene or the G418 (geneticin) resistance gene, P_{SV40} in the promoter from SV40.

Figure 3B shows the plasmid integrated into the genome.

Figure 4A is a schematic map of the pFrog-PCV.

Figure 4B is a schematic map of the vector pFrog-PCV as it would appear when integrated into the eukaryotic genome.

Figure 5 is the nucleotide sequence of HSV1-thymidine kinase gene fused to the zeocin resistance gene.

Figure 6A is the schematic map of pSOF-CMV.

Figure 6B is the schematic map of the vector pSOF-CMV once it is integrated into the eukaryotic genome.

Figure 7A is the schematic map of pSOF-PCV.

Figure 7B is the schematic map of the vector pSOF-PCV once it is integrated into the eukaryotic genome.

Figure 8A is a schematic map of the plasmid pSOF-IL6.

Figure 8B is a schematic map of the plasmid pSOF-IL6 once it is integrated into eukaryotic genomic DNA.

Figure 9A is a time chart of the procedure for validating the expression of the tk:sh ble gene in the pSOF-PCV transfectants.

Figure 9B is an autoradiograph of a Northern blot after it was probed with radiolabelled tk:sh ble.

Figure 9C is a graph showing the increase in the level of expression observed between the different transfectants after induction. (-) before induction (+) after induction.

Figure 10A is a time line of the method for determining the time course of induction of SOF-IL6.9 transfectants.

Figure 10 B is an autoradiograph of a northern blot of tk:sh ble RNA probed with the tk:sh ble probe showing the increase in expression in time after induction of SOF-IL6.9 transfectants.

Figure 10C is a graph of the increase in expression of tk:sh ble RNA after induction of SOF-IL6.9 transfectants.

Figure 11A is a time line of the method for determining the time course of induction of PCV-IL1.2 transfectants.

Figure 11B is an autoradiograph of a northern blot of tk:sh ble RNA probed with the tk:sh ble probe showing the increase in expression in time after induction of PCV-IL1.2 transfectants.

Figure 11C is a graph of the increase in expression of tk:sh ble RNA after induction of PCV-IL1.2 transfectants.

Figure 12A is the restriction map of pDOF-CMV.

Figure 12B shows the vector pDOF-CMV in transfectants after integration into the eukaryotic genome.

Figure 13A is a schematic map of pDOF-PCV.

Figure 13B is a schematic map of the vector pDOF-PCV once it is integrated into the eukaryotic genome.

Figure 14A is a schematic map of the plasmid pDOF-IL6.

Figure 14B is a schematic map of the configuration of the plasmid pDOF-IL6 once it is integrated into eukaryotic genomic DNA.

Figure 15A is a schematic map of pICOF-PCV.

Figure 15B is a schematic map of the pICOF-PCV vector in transfectants after integration into the eukaryotic genome.

Figure 16A is a schematic map of pICOF-CMV.

Figure 16B is a schematic map of the pICOF-CMV vector once it is integrated into the eukaryotic genome.

Figure 17A is a schematic map of the plasmid pICOF-IL6.

Figure 17B is a schematic map of the plasmid pICOF-IL6 once it is integrated into eukaryotic genomic DNA.

Figure 18 is a representation of the method by which the expression from a promoter is induced by the modulator resulting in expression of the survival gene.

Summary

The present invention recognizes that polynucleotides which encode proteins necessary for cell survival can be effectively used in living eukaryotic cells to functionally identify active portions of a genome directly or indirectly associated with a biological process. The present invention also recognizes that expression of such survival proteins can be selected in living cells incubated with a test chemical or modulator that directly or indirectly interacts with a portion of the genome having an integrated polynucleotide which encodes proteins necessary for cell survival such that transcription of the polynucleotide is induced or halted.

The present invention, thus, permits the rapid identification and isolation of genomic polynucleotides indirectly or directly associated with a defined biological

process and identification of compounds that modulate such processes and regions of the genome. Because the identification of active genomic polynucleotides is permitted in living cells, further functional characterization can be conducted using the same cells, and optionally, the same screening assay. The ability to functionally screen immediately after the rapid identification of a functionally active portion of a genome, without the necessity of transferring the identified portion of the genome into a secondary screening system, represents, among other things, a distinct advantage.

The invention provides for a method of identifying portions of a genome, e.g. genomic polynucleotides, in a living cell using a polynucleotide encoding a survival protein. Typically, the method involves inserting a polynucleotide encoding a survival protein into the genome of an organism using any method known in the art, developed in the future or described herein. Usually, the survival gene expression construct will be used to integrate a polynucleotide into a eukaryotic genome, as described herein. The cell, such as a eukaryotic cell, is usually contacted with a predetermined concentration of a modulator after integration of the survival polynucleotide. The presence of an active survival protein in the cell is usually then ascertained by placing the cell under selective cell growth pressures as described herein.

In one of its method aspects, the invention provides a method for identifying modulators that directly or indirectly modulate expression of a genomic polynucleotide comprising:

providing a nucleic acid sequence comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site integrated into a genomic polynucleotide in a eukaryotic genome contained in at least one living cell which survival polynucleotide is transcriptionally incompetent,

contacting said cell with a predetermined concentration of a modulator, and

placing the cell under survival conditions and identifying those cells which survive. Domain 1 of the survival polynucleotide is selected from the group consisting of the zeocin gene, hygromycin gene, neomycin gene, blasticidin S, puromycin gene. Domain 2 of the survival polynucleotide is selected from the group consisting of the thymidine kinase gene and the cytidine diaminase gene.

In another of its method aspects, the invention provides a method for identifying modulators, comprising:

(a) providing a nucleic acid sequence comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site and a known inducible promoter, which sequence is integrated into a eukaryotic genome contained in at least one living cell,

(b) contacting said cell with a predetermined concentration of a test chemical, and

(c) placing the cell under survival conditions and identifying those cells which survive.

This method may further comprise

(d) providing a nucleic acid sequence comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site and a known inducible promoter, which sequence is integrated into a eukaryotic genome contained in at least one living cell,

(e) contacting said cell with a predetermined concentration of a known modulator,

(f) placing the cell under survival conditions and identifying those cells which survive, and

(g) determining whether the percentage of cells that survive step (c) r is comparable to the percentage of cells that survive step (f).

In another method aspect, this invention provides a method for identifying intracellular pathways, comprising:

providing a plurality of eukaryotic cells, wherein the eukaryotic genome of each cell comprises a nucleic acid sequence comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site and a known inducible promoter, wherein said plurality of cells has a plurality of integration sites where said nucleic acid sequence has integrated,

contacting said plurality of eukaryotic cells with a modulator of interest,

placing the plurality of cells under survival conditions and identifying those cells which survive,

wherein survival of said cells indicates participation of said integration site in the intracellular pathway.

This invention further provides a method for identifying a promoter region capable of being modulated by a modulator, comprising:

providing a plurality of eukaryotic cells, wherein the eukaryotic genome of each cell comprises a nucleic acid sequence comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site, wherein said plurality of cells has a plurality of integration sites where said nucleic acid sequence has integrated,

contacting said plurality of eukaryotic cells with a modulator of interest,

placing the plurality of cells under survival conditions and identifying those cells which survive, and

isolating the promoter region at the integration site operably linked to the survival polynucleotide in the surviving cells.

This invention further provides a method for identifying an enhancer region capable of being modulated by a modulator, comprising:

providing a plurality of eukaryotic cells, wherein the eukaryotic genome of each cell comprises a nucleic acid sequence comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a known weak promoter region requiring an enhancer, a splice acceptor site and an internal ribosome entry binding site, wherein said plurality of cells has a plurality of integration sites where said nucleic acid sequence has integrated,

contacting said plurality of eukaryotic cells with a modulator of interest,

placing the plurality of cells under survival conditions and identifying those cells which survive, and

isolating the enhancer region operably linked to the survival polynucleotide in the surviving cells.

This invention further provides an ES cell comprising a nucleic acid sequence integrated into the genome of the cell comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site.

This invention further provides a plurality of ES cells each comprising a nucleic acid sequence integrated into the genome of the cell comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site wherein said plurality of cells has a plurality of integration sites where said nucleic acid sequence has integrated.

The invention also includes powerful methods and compositions for identifying physiologically relevant cellular pathways and proteins of interest of known, unknown or partially known function. As shown in FIG. 18 a pathway

may have more than one major intracellular signal. Two major intracellular pathways are shown ("A" and "B"). Each intracellular signal pathway may also have multiple branches. Each arm is shown as having three signaling pathways (Al, Al, and A3, and B1, B2, and B3). By generating a library of clones with a survival gene expression construct, genomic polynucleotides for each signal pathway can be tagged or reported by placing the cells under selective conditions. Pathways not effected by the modulator (shown as Cl, C2, and C3) are also tagged with survival gene expression construct. Because the modulator only modulates the expression of pathways Al, A2, A3, BI, B2, and B3, only clones corresponding to these genomic integration sites are identified as being responsive to the modulator. Clones corresponding to sites Cl, C2, and C3 remain unaltered and are not responsive to the modulator. Any individual, modulated clone can be immediately isolated, if not already isolated, and used for a drug discovery assay to screen test chemicals for activity for modulating the reported pathway, as described herein.

The invention also includes tools for pathway identification and drug discovery that can be applied to a number of targets of interest and therapeutic areas including, proteins of interest, physiological responses even in the absence of a definitive target (e.g. immune response, signal transduction, neuronal function and endocrine function), viral targets, and orphan proteins.

Detailed Description of the Invention

Definitions

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Generally, the nomenclature used herein and the laboratory procedures in cell culture, molecular genetics, and nucleic acid

chemistry and hybridization described below are those well known and commonly employed in the art. Standard techniques are used for recombinant nucleic acid methods, polynucleotide synthesis, and microbial culture and transformation (e.g., electroporation. lipofection). Generally, enzymatic reactions and purification steps are performed according to the manufacturers' specifications. The techniques and procedures are generally performed according to conventional methods in the art and various general references (see generally, Sambrook et al. *Molecular Cloning A Laboratory Manual*, 2d ed. (1989) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. and Ausubel et al., (1989) "Current Protocols in Molecular Biology" John Wiley & Sons, Baltimore MA, which are incorporated herein by reference) which are provided throughout this document. The nomenclature used herein and the laboratory procedures in analytical chemistry, organic synthetic chemistry, and pharmaceutical formulation described below are those well known and commonly employed in the art. Standard techniques are used for chemical syntheses, chemical analysis, pharmaceutical formulation and delivery. As employed throughout the disclosure, the following terms, unless otherwise indicated, shall be understood to have the following meanings:

"Isolated polynucleotide" refers to a polynucleotide of genomic, cDNA, or synthetic origin or some combination thereof, which by virtue of its origin the "isolated polynucleotide" (1) is not associated with the cell in which the "isolated polynucleotide" is found in nature, or (2) is operably linked to a polynucleotide to which it is not linked in nature.

"Isolated protein" refers to a protein encoded by DNA, cDNA, recombinant RNA, or synthetic origin or some combination thereof, which by virtue of its origin the "isolated protein" (1) is not associated with proteins with which it is normally found in nature, or (2) is isolated from the cell in which it normally occurs or (3) is isolated free of other proteins from the same cellular

source, e.g. free of human proteins, or (4) is expressed by a cell from a different species, or (5) does not occur in nature.

"Polypeptide" as used herein is a generic term to refer to native protein, fragments, or analogs of a polypeptide sequence. Hence, native protein, fragments, and analogs are species of the polypeptide genus.

Naturally-occurring" as used herein, as applied to an object, refers to the fact that an object can be found in nature. For example, a polypeptide or polynucleotide sequence that is present in an organism (including viruses) that can be isolated from a source in nature and which has not been intentionally modified by man in the laboratory is naturally-occurring.

"Operably linked" refers to a juxtaposition wherein the components so described are in a relationship permitting them to function in their intended manner. A control sequence "operably linked" to a coding sequence is ligated in such a way that expression of the coding sequence is achieved under conditions compatible with the control sequences.

"Control sequence" refers to polynucleotide sequences which are necessary to effect the expression of coding and non-coding sequences to which they are ligated. The nature of such control sequences differs depending upon the host organism; in prokaryotes, such control sequences generally include promoter, ribosomal binding site, and transcription termination sequence; in eukaryotes, generally, such control sequences include promoters, enhancers and transcription termination sequence. The term "control sequence" is intended to include, at a minimum, components whose presence can influence expression, and can also include additional components whose presence is advantageous, for example, leader sequences and fusion partner sequences.

The "promoter region" refers to the control sequence located 5' to the gene which controls the transcription of the gene. The term "promoter region" is intended to include, at a minimum components which presence can influence expression such as promoter sequences and can also include additional components which presence is advantageous.

The "enhancer region" refers to the cis-acting control sequence which may be located either 5' or 3' of the expressed gene. It may also be located within introns and the coding region itself. It influences transcription of the gene.

The term "transcriptionally incompetent" means that a polynucleotide or gene lacks a promoter region and thus is incapable of being transcribed. A "transcriptionally incompetent gene" may be transcribed if it is integrated into an active region of the genome such that the gene is operably-linked to an active promoter region and/or enhancer region.

"Polynucleotide" refers to a polymeric form of nucleotides of at least 10 bases in length, either ribonucleotides or deoxynucleotides or a modified form of either type of nucleotide. The term includes single and double stranded forms of DNA. "Genomic polynucleotide" refers to a portion of a genome. "Active genomic polynucleotide" or "active portion of a genome" refer to regions of a genome that can be up regulated, down- regulated or both, either directly or indirectly, by a biological process. "Directly," in the context of a biological process or processes, refers to direct causation of a process that does not require intermediate steps, usually caused by one molecule contacting or binding to another molecule (the same type or different type of molecule). For example, molecule A contacts molecule B which causes molecule B to exert effect X that is part of a biological process. "Indirectly," in the context of a biological process or processes, refers to indirect causation that requires intermediate steps, usually

caused by two or more direct steps. For example, molecule A contacts molecule B to exert effect X which in turn causes effect Y.

"Survival polynucleotide which encodes survival proteins" or "survival gene" refers to a polynucleotide encoding a protein which contains two domains, Domain 1 when expressed enables the cells to survive in the presence of a selection compound; whereas domain 2 enables cell survival only when domain 2 is not expressed in the presence of a different selection compound.

Some examples of proteins with the characteristics of domain 1 are markers that can be selected positively by their ability to induce either zeocin (zeo), hygromycin (Hyg), neomycin (neo), puromycin (PAC), or blasticidin S (BlaS) resistance in cells. Some examples of proteins with the chraracteristics of domain 2 are markers with negative selectabilities based on the thymidine kinase (tk) gene of *Herpes simplex* virus (HSV) or the cytidine deaminase (codA) gene of *E. coli* (Karreman C. "A new set of positive/negative selectable markers for mammalian cells" *Gene* 1998:218:57-61).

Some examples of survival proteins having both a domain 1 and a domain 2 are fusion proteins of:

1. thymidine kinase (tk gene of *Herpes simplex* virus (HSV) and basticidin S deaminase (bsd);
2. thymidine kinase (tk) gene of *Herpes simplex* virus (HSV) and *Streptoalloteichus hindustanus* bleomycin-resistance gene product (Sh ble); and
3. basticidin S deaminase (bsd) and cytidine deaminase (codA) fused to uracil phosphoribosyltransferase.

The sensitivities of certain proteins with the characteristics of domain 2 may be further enhanced by fusing to yet another enzyme (e.g., codA::upp). The

transfection of mammalian cells with a construct containing the fused genes from a non-eukaryote allows the prodrug 5-fluorocytosine (5-FC) to be used efficiently in negative selection strategies to generate 5FdUMP, an irreversible inhibitor of TTP production.

The term "survival conditions" means those conditions sufficient to kill greater than 50% of the cells not expressing the survival protein, preferably the survival conditions are those sufficient to kill greater that 75% of the cells not expressing the survival protein and most preferably the conditions are such that greater that 90% of the cells not expressing the survival protein are killed. Cell death can be determined by various methods known in the art, including failure to observe cell colony growth under survival conditions. For example, and without being limiting, if a population of cells, some having the zeocin resistance gene integrated into the cell genome where it is constitutively expressed, are placed under survival conditions (i.e. in the presence of zeocin) then those cells constitutively expressing the zeocin resistance gene will continue to replicate and form colonies and those cells not expressing the gene will die.

"Puromycin polynucleotide" refers to a polynucleotide encoding a protein with puromycin-N-acetyltransferase activity. (de la Luna et al., Gene Vol 62, 1988 121-126)

"Thymidine kinase polynucleotide" refers to a polynucleotide encoding a protein with thymidine kinase (tk) activity. Thymidine kinase catalyzes the conversion of thymidine to deoxythimidine-monophosphate. The tk gene is preferably the one derived from herpes simplex virus (HSV).

Herpes Simplex virus (HSV) produces a thymidine kinase that can convert ganciclovir (GCV) to ganciclovir-monophosphate (GCV-MP). Expression of

HSVtk sensitizes transfected cells to ganciclovir. GCV is a guanosine analog that can be phosphorylated by HSV tk to GCV-MP. GCV-MP is then converted to the diphosphate and triphosphate forms by endogenous kinases. GCV-triphosphate lacks the 3' OH on the doxyribose as well as the bond between the 2' and 3' carbons which are necessary for DNA chain elongation. As a result, GCV-triphosphate integration causes prmature DNA chain termination and leads to apoptosis. Hence cells that express HSV tk do not survive in the presence of GCV selection, whereas cells that express no or low level of HSV-tk can survive in the presence of GCV selection.

"Kanamycin or neomycin polynucleotide" refers to a polynucleotide that confers resistance to kanamycin. Two genes (kan) from transposon 5 and transposon 601, respectively that encode aminoglycoside-3'-phosphotransferases (APH) I and II have been isolated. These enzymes phosphorylate antibiotics such as kanamycin, neomycin or related aminoglycoside compounds such as G418 (geneticin) and inactivate them.

"Blasticidin S polynucleotide" refers to a polynucleotide encoding a protein that confers blasticidin resistance to cells through its basticidin S deaminase activity. Expression vectors encoding blasticidin or blasticidin fusion proteins have been described (Karreman C. "A new set of positive/negative selectable markers for mammalian cells" *Gene* 1998 Sep 18, 218(1-2): 57-61).

Zeocin polynucleotide" refers to a polynucleotide encoding a protein that confers resistance to the antibiotic zeocin, a versatile antibiotic that is used for selection in mammalian cells, plants, yeasts and bacteria. Zeocin acts by binding to DNA and cleaving it, causing cell death. An example is the product of the Sh ble (Streptoalloteichus hindustanus bleomycin-resistance) gene that confers resistance to zeocin and zeocell. This 14 kD protein stoichiometrically binds to

zeocin and inhibits DNA cleavage and subsequence cell death (Invivogen, San Diego CA).

"Sequence homology" refers to the proportion of base matches between two nucleic acid sequences or the proportion amino acid matches between two amino acid sequences. When sequence homology is expressed as a percentage, e.g., 50%, the percentage denotes the proportion of matches over the length from a desired sequence that is compared to some other sequence. Gaps (in either of the two sequences) are permitted to maximize matching; gap lengths of 15 bases or less are usually used, 6 bases or less are preferred with 2 bases or less more preferred. When using oligonucleotides as probes or modulators the sequence homology between the target nucleic acid and the oligonucleotide sequence is generally not less than 17 target base matches out of 20 possible oligonucleotide base pair matches (85%); preferably not less than 9 matches out of 10 possible base pair matches (90%), and most preferably not less than 19 matches out of 20 possible base pair matches (95%).

"Selectively hybridize" refers to the ability to detectably and specifically bind. Polynucleotides, oligonucleotides and fragments thereof selectively hybridize to target nucleic acid strands under hybridization and wash conditions that minimize appreciable amounts of detectable binding to nonspecific nucleic acids. High stringency conditions can be used to achieve selective hybridization conditions as known in the art and discussed herein. Generally, the nucleic acid sequence homology between the polynucleotides, oligonucleotides, and fragments thereof and a nucleic acid sequence of interest will be at least 30%, and more typically with preferably increasing homologies of at least about 40%, 50%, 60%, 70%, and 90%.

Typically, hybridization and washing conditions are performed at high stringency according to conventional hybridization procedures. Positive clones are isolated and sequenced. For illustration and not for limitation, a full-length polynucleotide may be labeled and used as a hybridization probe to isolate genomic clones from the appropriate target library in λEMBL4 or λGEM1 1 (Promega Corporation, Madison, Wisconsin); typical hybridization conditions for screening plaque lifts (Benton and Davis (1978) Science 196:180) can be: 50% formamide, 5 x SSC or SSPE, 1-5 x Denhardt's solution, 0.1-1% SDS, 100-200 μg sheared heterologous DNA or tRNA, 0-10% dextran sulfate, I x10^5 to 1 x 10^7 cpm/ml of denatured probe with a specific activity of about I x 10^5 cpm/μg, and incubation at 42°C for about 6-36 hours. Prehybridization conditions are essentially identical except that probe is not included and incubation time is typically reduced. Washing conditions are typically 1-3 x SSC, 0.1-1% SDS, 50-70°C with change of wash solution at about 5-30 minutes. Cognate sequences, including allelic sequences, can be obtained in this manner.

Two amino acid sequences are homologous if there is a partial or complete identity between their sequences. For example, 85% homology means that 85% of the amino acids are identical when the two sequences are aligned for maximum matching. Gaps (in either of the two sequences being matched) are allowed in maximizing matching; gap lengths of 5 or less are preferred with 2 or less being more preferred. Alternatively and preferably, two protein sequences (or polypeptide sequences derived from them of at least 30 amino acids in length) are homologous. As this term is used herein, if they have an alignment score of at more than 5 (in standard deviation units) *using* the program ALIGN with the mutation data matrix and a gap penalty of 6 or greater. See Dayhoff, M.O.. in *Atlas of Protein Sequence and Structure*, (1972), volume 5, National Biomedical Research Foundation, pp 101-110, and Supplement 2 to this volume, pp.110. The two sequences or parts thereof are more preferably homologous if their amino

acids are greater than or equal to 30% identical when optimally aligned using the ALIGN program.

"Corresponds to" refers to a polynucleotide sequence that is homologous (i.e., is identical, not strictly evolutionarily related) to all or a portion of a reference polynucleotide sequence, or to a polypeptide sequence that is identical to all or a portion of a reference polypeptide sequence. In contradistinction, the term "complementary to" is used herein to mean that the polynucleotide sequence is homologous to all or a portion of the complement of a reference polynucleotide sequence. For illustration, the nucleotide sequence "TATAC" corresponds to a reference sequence "TATAC" and is complementary to a reference sequence "GTATA"

The following terms are used to describe the sequence relationships between two or more polynucleotides: "reference sequence," "comparison window," "sequence identity," "percentage of sequence identity," and "substantial identity." A "reference sequence" is a defined sequence used as a basis for a sequence comparison; a reference sequence may be a subset of a larger sequence, for example, as a segment of a full-length cDNA or gene sequence given in a sequence listing or may comprise a complete cDNA or gene sequence. Generally, a reference sequence is at least 20 nucleotides in length, frequently at least 25 nucleotides in length, and often at least 50 nucleotides in length. Since two polynucleotides may each (1) comprise a sequence (i.e., a portion of the complete polynucleotide sequence) that is similar between the two polynucleotides, and (2) may further comprise a sequence that is divergent between the two polynucleotides. sequence, comparisons between two (or more) polynucleotides are typically performed by comparing sequences of the two polynucleotides over a "comparison window" to identify and compare local regions of sequence similarity.

A "comparison window", as used herein, refers to a conceptual segment of at least 20 contiguous nucleotide positions wherein a polynucleotide sequence may be compared to a reference sequence of at least 20 contiguous nucleotides and wherein the portion of the polynucleotide sequence in the comparison window may comprise additions or deletions (i.e., gaps) of 20 percent or less as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. Optimal alignment of sequences for aligning a comparison window may be conducted by the local homology algorithm of Smith and Waterman (1981) *Adv. Appl. Math.* 2:482, by the homology alignment algorithm of Needleman and Wunsch (1970) *J. Mol. Biol.* 48: 443, by the search for similarity method of Pearson and Lipman (1988) *Proc. Natl. Acad. Sci.* (U.S.A.) 85: 2444, by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package Release 7.0, Genetics Computer Group, 573 Science Dr., Madison, WI), or by inspection, and the best alignment (i.e., resulting in the highest percentage of homology over the comparison window) generated by the various methods is selected.

The term "sequence identity" means that two polynucleotide sequences are identical (i.e., on a nucleotide-by-nucleotide basis) over the window of comparison. The term "percentage of sequence identity" is calculated by comparing two optimally aligned sequences over the window of comparison, determining the number of positions at which the identical nucleic acid base (e.g., A, T, C, G, U, or I) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison (i.e., the window size), and multiplying the result by 100 to yield the percentage of sequence identity.

The terms "substantial identity" as used herein denotes a characteristic of a polynucleotide sequence, wherein the polynucleotide comprises a sequence that has at least 30 percent sequence identity, preferably at least 50 to 60 percent sequence identity, more usually at least 60 percent sequence identity as compared to a reference sequence over a comparison window of at least 20 nucleotide positions, frequently over a window of at least 25-50 nucleotides, wherein the percentage of sequence identity is calculated by comparing the reference sequence to the polynucleotide sequence which may include deletions or additions which total 20 percent or less of the reference sequence over the window of comparison.

As applied to polypeptides, the term "substantial identity" means that two peptide sequences, when optimally aligned, such as by the programs GAP or BESTFIT using default gap weights, share at least 30 percent sequence identity, preferably at least 40 percent sequence identity, more preferably at least 30 percent sequence identity, and most preferably at least 60 percent sequence identity. Preferably, residue positions which are not identical, differ by conservative amino acid substitutions. Conservative amino acid substitutions refer to the interchangeability of residues having similar side chains. For example, a group of amino acids having aliphatic side chains is glycine, alanine, valine, leucine, and isoleucine; a group of amino acids having aliphatic-hydroxyl side chains is serine and threonine; a group of amino acids having amide-containing side chains is asparagine and glutamine; a group of amino acids having aromatic side chains is phenylalanine, tyrosine, and tryptophan; a group of amino acids having basic side chains is lysine, arginine, and histidine; and a group of amino acids having sulfur-containing side chains is cysteine and methionine. Preferred conservative amino acids substitution groups are: valine-leucine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, glutamic-aspartic, and asparagine-glutamine.

The terms "substantial identity" as used herein denotes a characteristic of a polynucleotide sequence, wherein the polynucleotide comprises a sequence that has at least 30 percent sequence identity, preferably at least 50 to 60 percent sequence identity, more usually at least 60 percent sequence identity as compared to a reference sequence over a comparison window of at least 20 nucleotide positions, frequently over a window of at least 25-50 nucleotides, wherein the percentage of sequence identity is calculated by comparing the reference sequence to the polynucleotide sequence which may include deletions or additions which total 20 percent or less of the reference sequence over the window of comparison.

As applied to polypeptides, the term "substantial identity" means that two peptide sequences, when optimally aligned, such as by the programs GAP or BESTFIT using default gap weights, share at least 30 percent sequence identity, preferably at least 40 percent sequence identity, more preferably at least 30 percent sequence identity, and most preferably at least 60 percent sequence identity. Preferably, residue positions. which are not identical, differ by conservative amino acid substitutions. Conservative amino acid substitutions refer to the interchangeability of residues having similar side chains. For example, a group of amino acids having aliphatic side chains is glycine, alanine, valine, leucine, and isoleucine; a group of amino acids having aliphatic-hydroxyl side chains is serine and threonine; a group of amino acids having amide-containing side chains is asparagine and glutamine; a group of amino acids having aromatic side chains is phenylalanine, tyrosine, and tryptophan; a group of amino acids having basic side chains is lysine, arginine, and histidine; and a group of amino acids having sulfur-containing side chains is cysteine and methionine. Preferred conservative amino acids substitution groups are: valine-leucine-isoleucine, phenylalanine-tyrosine, lysine-arginine, alanine-valine, glutamic-aspartic, and asparagine-glutamine.

"Polypeptide fragment" refers to a polypeptide that has an amino-terminal and/or carboxy-terminal deletion but where the remaining amino acid sequence is usually identical to the corresponding positions in the naturally-occurring sequence deduced, for example, from a full-length cDNA sequence. Typically, analog polypeptides comprise a conservative amino acid substitution (or addition or deletion) with respect to the naturally occurring sequence. Analogs typically are at least 300 amino acids long, preferably at least 500 amino acids long or longer, most usually being as long as full-length naturally-occurring polypeptide.

"Modulation" refers to the capacity to change a biological activity or process (e.g., to enhance or inhibit enzyme activity or receptor binding activity). Such enhancement or inhibition maybe contingent on the occurrence of a specific event, such as activation of a signal transduction pathway, and/or may be manifest only in particular cell types.

The term "modulator" refers to a chemical substance, extract that is capable of modulation as defined above. A modulator may be macromolecular or molecular in nature, naturally occurring or otherwise obtained through chemical or biological synthesis, or a combination of these, and may be a purified single biochemical substance or a mixture or extract of substances from a biological organism or cells.

The term "test chemical" refers to a chemical to be tested by one or more method(s) of the invention for modulatory activity. Usually, various predetermined concentrations of test chemicals are used for screening such as 0.01 μM, 0.1 μM, 1.0 μM, and 10.0 μM.

The term "target" refers to a biochemical entity involved in a biological process. Targets may be biological macromolecules that play a useful role in the

physiology or biology of an organism. A therapeutic chemical binds to a target to alter or modulate its function. As used herein, targets can include cell surface receptors, C-proteins, kinases, ion channels, phopholipases and other proteins mentioned herein as well as nucleic acid sequences or structures.

The terms "label" or "labeled" refers to incorporation of a detectable marker, e.g., by incorporation of a radiolabeled amino acid or attachment to a polypeptide of biotinyl moieties that can be detected by marked avidin (e.g., streptavidin containing a fluorescent marker or enzymatic activity that can be detected by optical or colorimetric methods). Various methods of labeling polypeptides and glycoproteins are known in the art and may be used. Examples of labels for polypeptides include, but are not limited to, the following: radioisotopes (e.g., ^{3}H, ^{14}C, ^{35}S, $^{125}1$, ^{131}I), fluorescent labels (e.g., FITC, rhodamine. and lanthanide phosphors), enzymatic labels (or reporter genes) (e.g., enzymatic reporter genes horseradish peroxidase, β-galactosidase, luciferase and alkaline phosphatase; and non-enzymatic reporter genes (e.g.. fluorescent proteins)), chemiluminescent, biotinyl groups, predetermined polypeptide epitopes recognized by a secondary reporter (e.g., leucine zipper pair sequences, binding sites for secondary antibodies, metal binding domains, epitope tags).

"Substantially pure" refers to an object species that is the predominant species present (i.e., on a molar basis it is more abundant than any other individual species in the composition), and preferably a substantially purified fraction is a composition wherein the object species comprises at least about 50 percent (on a molar basis) of all macromolecular species present. Generally, a substantially pure composition will comprise more than about 80 percent of all macromolecular species present in the composition, more preferably more than about 85%, 90%, 95%, and 99%. Most preferably, the object species is purified to essential homogeneity (contaminant species cannot be detected in the

composition by conventional detection methods) wherein the composition consists essentially of a single macromolecular species.

"Pharmaceutical agent or drug" refers to a chemical or composition capable of inducing a desired therapeutic effect when properly administered (e.g. using the proper amount and delivery modality) to a patient.

Other chemistry terms herein are used according to conventional usage in the art, as exemplified by The McGraw-Hill Dictionary of Chemical Terms (ed. Parker, S., 1985), McGraw-Hill, San Francisco, incorporated herein by reference).

Introduction

The present invention recognizes that survival polynucleotides can be effectively used in living eukaryotic cells to functionally identity active portions of a genome directly or indirectly associated with a biological process. The present invention also recognizes that the expression of the survival protein can be measured by placing the cells under survival conditions. The present invention, thus, permits the rapid identification and isolation of genomic polynucleotides indirectly or directly associated with a defined biological process and identification of compounds that modulate such processes and regions of the genome. Because the identification of active genomic polynucleotides is permitted in living cells, further functional characterization can be conducted using the same cells, and optionally, the same screening assay. The ability to functionally screen immediately after the rapid identification of a functionally active portion of a genome, without the necessity of transferring the identified portion of the genome into a secondary screening system, represents, among other things, a distinct

advantage over an application of a prior art reporter gene and methods described herein.

As a non-limiting introduction to the breadth of the invention, the invention includes several general and useful aspects, including:

1) a method for identifying genes or gene products directly or indirectly associated with a biological process of interest (that can be modulated by a compound) by operably linking a genomic polynucleotide to a polynucleotide encoding a survival protein.

2) a method for identifying modulators (e.g. orphan proteins or known proteins) or compounds that directly or indirectly modulate transcription by operably linking a genomic polynucleotide to a polynucleotide encoding a survival protein,

3) a method of screening for an active genomic polynucleotide (e.g. enhancer, promoter or coding region in the genome) that is directly or indirectly associated with a modulatable biological process of interest by operably linking a genomic polynucleotide to a survival polynucleotide,

4) polynucleotides related to the above methods, and

5) ES cells transformed with the polynucleotides of the present invention.

These aspects of the invention, as well as others described herein, can be achieved by using the methods and compositions of matter described herein. To gain a full appreciation of the scope of the invention, it will be further recognized that various aspects of the invention can be combined to make desirable

embodiments of the invention. For example, the invention includes a method of identifying compounds that modulate active genomic polynucleotides operably linked to a survival protein. Such combinations result in particularly useful and robust embodiments of the invention.

Methods for Rapidly Identifying Functional Portions of a Genome

The invention provides for a method of identifying portions of a genome, e.g. genomic polynucleotides, in a living cell using a polynucleotide encoding a survival protein. Typically, the method involves inserting a polynucleotide encoding a survival protein into the genome of an organism using any method known in the art, developed in the future or described herein. Usually, an expression construct will be used to integrate a survival polynucleotide into a eukaryotic genome, as described herein. The cell, such as a eukaryotic cell, is usually contacted with a predetermined concentration of a modulator after integration of the survival polynucleotide and the cell is placed under survival conditions.

Once the survival polynucleotides are integrated into the genome of interest, they come under the transciptional control of the genome of the host cell. Integration into the genome is usually stable, as described herein and known in the art. Transcriptional control of the genome often results from receptor (e.g. intracellular or cell surface receptor) activation, which can regulate transcriptional and translational events to change the amount of protein present in the cell.

Vectors and Integration

Vectors, such as viral and plasmid vectors, can be used to introduce genes or genetic material of the invention into cells, preferably by integration into the

host cell genome. Such viral vectors can be any appropriate viruses, such as retroviruses, adenoviruses, adeno-associated viruses, papillomaviruses, herpes viruses, or any ecotropic or amphitropic virus, preferably a retrovirus. The viruses can be, for example, retroviruses or any other virus modified to be replicatively deficient, cytomegalovirus, Friend leukemia virus, SIV, HIV, Rous Sarcoma Virus, or Maloney virus such as Moloney murine leukemia virus.

Vectors, such as retrovirus vectors, can also encode an operable selective protein so that cells that have been transformed can be positively selected for. Such selective genes would be transcriptionally competent having an active control region. Such positive selection proteins necessarily would not be the same as those encoded within the survival polynucleotide. Such selective proteins can be antibiotic resistance factors, such as neomycin resistance, such as *NEO*. Alternatively, cells can be negatively selected for using an enzyme, such as herpes simplex virus thymidine kinase (HSVTK) that transforms a non-cytotoxic prodrug into a cytotoxic drug. Viral vectors, such as retroviral vectors, are available that are suitable for these purposes, such as PSIR vector (available from ClonTech of California with PT67 packaging cells) GgU3Hisen and GgTNKneoU3 and GgTKNeoen variants of Moloney murine leukemia virus, are available. Vector modifications can be made that allow more efficient integration into the host cell genome. Such modifications include sequences that enhance integration or known methods to promote nucleic acid transportation into the nucleus of the host cell. Retro-viral vectors such as those described in U.S. Patent Number 5,364,783 by Ruley and von Melchner can also be used to increase transfection efficiency.

Vectors can also be used with liposomes or other vesicles that can transport genetic material into a cell. Appropriate structures are known in the art. The liposomes can include vectors such as plasmids or yeast artificial chromosomes (YACs), which can include genetic material to be introduced into the cell.

Plasmids can also be introduced into cells by any known methods, such as electroporation, calcium phosphate, or lipofection. DNA fragments, without a plasmid or viral vector can also be used.

Survival polynucleotides can be placed on a variety of plasmids for integration into a genome and to identify genes from a large variety of organisms. Standard techniques are used to introduce these polynucleotides into a cell or whole organism (e.g., as described in Sambrook, S., Fritsch, E.F. and Maniatis, T. Expression of cloned genes in cultured mammalian cells. In: Molecular Cloning, edited by Nolan, C. New York: Cold Spring Harbor Laboratory Press, 1989). Resistance markers can be used to select for successfully transfected cells.

If a survival polynucleotide expression construct is selected for integrating a survival polynucleotide into a eukaryotic genome, it will usually contain at least a survival polynucleotide operably linked to a splice acceptor and optionally a splice donor. Alternatively, the survival polynucleotide may be operably linked to any means for integrating a polynucleotide into a genome, preferably for integration into an intron of a gene to produce an in frame translation product. The survival polynucleotide expression construct can optionally comprise, depending on the application, an IRES element, a poly A site, translational start site (e.g. a Kozak sequence) an LTR (long terminal repeat) and a selectable marker.

Sequences for Assisting Integration

The survival polynucleotide expression construct typically includes sequences for integration, especially sequences designed to target or enhance integration into the genome. A splice acceptor site can be operably linked to the survival polynucleotide to facilitate expression upon integration into an intron. Usually, a fusion RNA will be created with the coding region of an adjacent

operably linked portion of the exon. A splice acceptor sequence is a sequence at the 3' end of an intron where it junctions with an exon. The consensus sequences for a splice acceptor sequence is usually made up of a pyrmidine rich region preceding the dinucleotide AG located at the 3' end of an intron.

A splice donor site may be operably linked to the survival polynucleotide to facilitate integration in an intron to promote expression by requiring a poly-adenylation sequence.

As an alternative to a splice donor site, a poly A site may be operably linked to the survival polynucleotide. Poly-adenylation signals, i.e poly A sites, include SV40 poly A sites, such as those described in the Invitrogen Catalog 1996 (California). In some instances, it may be desirable to include in the survival expression construct a translational start site. For instance, a translational start site allows for survival protein expression even if the integration occurs in non-coding regions. Usually, such sequences *will* not reduce the expression of a highly expressed gene. Translational start sites include a "Kozak sequence" and are the preferred sequences for expression in mammalian cells described in Kozak, M., *J. Cell Biol* 108:229-241(1989).

It is also preferable, when using mammalian cells, to include an IRES ("internal ribosome entry binding site") element in the survival polynucleotide expression construct. Typically, an IRES element will improve the yield of expressing clones. One caveat of integration vectors is that only one in three insertions into an intron will be in frame and produce a functional reporter protein. This limitation can be reduced by cloning an IRES sequence between the splice acceptor site and the reporter gene (e.g., a survival polynucleotide). This eliminates reading frame restrictions and possible functional inactivation of the reporter protein by fusion to an endogenous protein. IRES elements include those

from piconaviruses, picorna-related viruses, and hepatitis A and C. Preferably. the IRES element is from a poliovirus. Specific IRES elements can be found, for instance, in W09611211 by Das and Coward published 4/16/96, EP 585983 by Zurr published 3/7/96, W09601 324 by Berlioz published 1/18/96 and W09424301 by Smith published October 27, 1994, all of which are herein incorporated by reference.

To improve selection of survival polynucleotide into a genome, a selectable marker can be used in the survival expression construct. Selectable markers for mammalian cells are known in the art, and include for example, thymidine kinase, dihydrofolate reductase (together with methotrexate as a DHFR amplifier), aminoglycoside phosphotransferase, hygromycin B phosphotransferase, asparagine synthetase, adenosine deaminase, metallothionien, and antibiotic resistant genes such as neomycin resistance gene. Selectable markers for non-mammalian cells are known in the art and include genes providing resistance to antibiotics, such as kanamycin, tetracycline, and ampicillin.

The invention can be readily practiced with genomes having intron/exon structures. Such genomes include those of mammals (e.g., human, rabbit, mouse, rat, monkey, pig and cow), vertebrates, insects and yeast. Intron-targeted vectors are more commonly used in mammalian cells as introns (intervening sequences) are considerably larger than exons (mRNA coding regions) in mammals. Intron targeting can be achieved by cloning a splice acceptor or 3' intronic sequences upstream of a survival polynucleotide gene followed by a polyadenylation signal or 5' intronic splice donor site. When the vector inserts into an intron, the reporter gene is expressed under the same control as the gene into which it has inserted.

The invention can also be practiced with genomes having reduced numbers of, or lacking, intron/exon structures. For lower eukaryotes, which have simple

genomic organization, i.e. containing few and small introns, exon-targeted vectors can be used. Such vectors include survival polynucleotides operably linked to a poly-adenylation sequence and optionally to an IRES element. Lower eukaryotes include yeast, and fungi and pathogenic eurokaryotes (e.g. parasites and microoganisms). For genomes lacking intron/exon structures restriction enzyme integration, transposon induced integration or selection integration can be used for genomic integration. Such methods include those described by Kuspa and Loomis, PNAS 89: 8803-8807(1992) and Derbyshire, K.M., Gene Nov. 7.143-144(1995). Retroviral vectors can also be used to integrate survival polynucleotides into a genome (e.g., eukaryotic), such as those methods and composition described in U.S. Patent Number 5,364,783.

Typically, integration will occur in the regions of the genome that are accessible to the integration vector. Such regions are usually active portions of the genome where there is increased genome regulatory activity, e.g. increased polymerase activity or a change in DNA binding by proteins that regulate transcription of the genome. Many embodiments of the invention described herein can result in random integration, especially in actively transcribed regions.

Integration into Active Portions of the Genome

Integration, however, can be directed to regions of the genome active during specific types of genome activity. For instance, integration at sites in the genome that are active during specific phases of the cell cycle can be promoted by synchronizing the cells in a desired phase of the cell cycle. Such cell cycle methods include those known in the art, such as serum deprivation or alpha factors (for yeast). Integration may also be directed to regions of the genome active during cell regulation by a chemical, such as an antagonist or agonist for a receptor or some other chemical that increase or decreases or otherwise modulates

genome activity. By adding the chemical of interest, genome activity can be increased often in specific regions to promote integration of an integration vector (e.g. as a reporter gene construct), including those of the invention, into such regions of the genome.

For instance, a nuclear receptor activator (general or specific) could be applied to activate the cells prior or during integration in order to promote integration of reporter genes at sites in the genome that become more active during nuclear receptor activation. Such cells could then be screened with the same or different nuclear receptor activator to identify which clones, and which portions of the genome are active during nuclear receptor activation. Any agonists, antagonists and modulators of the receptors described herein can be used in such a manner, as well as any other chemicals that increase or decrease genome activity.

Cells for Integration into the Genome

The cells used in the invention will typically correspond to the genome of interest. For example, if regions of the human genome are desired to be identified, then human cells containing a proper genetic complement will generally be used. Libraries, however, could be biased by using cells that contain extra-copies of certain chromosomes or other portions of the genome. Cells that do not correspond to the genome of interest can also be used if the genome of interest or significant portions of the genome of interest can be replicated in the cells, such as making a human-mouse hybrid.

Additionally, by the appropriate choice of cells and expressed proteins, identification and screening assays can be constructed that detect active portions of the genome associated with a biological process that requires, in whole or part, the presence of a particular protein (protein of interest). Cells can be selected

depending on the type of proteins that are expressed (homologously or heterologously) or from the type of tissue from which the cell line or explant was originally generated. If the identification of portions of the genome activated by a particular type of protein is desired, then the cell used should express that protein.

The cells express a protein homologously, i.e. expression of the desired protein normally or naturally occurs in the cells. Alternatively, the cells may be directed to express a protein heterologously, i.e. expression of a desired protein which does not normally or naturally occur in the cells. Such heterologous expression can be directed by "turning on" the gene in the cell encoding the desired protein or by transfecting the cell with a polynucleotide encoding the desired protein (either by constitutive expression or inducible expression). Inducible expression is preferred if it is thought that the expressed protein of interest may be toxic to the cells.

Many cells can be used with the invention. Such cells include, but are not limited to adult, fetal, or embryonic cells. These cells can be derived from the mesoderm, ectoderm, or endoderm and can be stem cells, such as embryonic or adult stem cells, or adult precursor cells. The cells can be of any lineage, such as vascular, neural, cardiac, fibroblasts, lymphocytes, hepatocytes, cardiac, hematopocitic, pancreatic, epidermal, myoblasts, or myocytes. Other cells include baby hamster kidney (BHK) cells (ATCC No. CCL1O)1 mouse L cells (ATCC No. CCLI.3), Jurkats (ATCC No. TB 152) and 153 DG44 cells (see, Chasm (1986) *Cell. Molec. Genet.* 12:555) human embryonic kidney (HEK) cells (ATCC No. CRL1S73), Chinese hamster ovary (CHO) cells (ATCC Nos. CRL96IS, CCL6I, CRL9096), PCI2 cells (ATCC No CRLI 7.21) and COS-7 cells (ATCC No. CRLI 651). Preferred cells include mouse embryonic stem cells, Jurkat cells, CHO cells, neuroblastoma cells, P19 cells, Fl I cells, NT-2 cells and REK 293 cells, such as those described in U.S. Patent No.5,024,939 and by Stiliman et al.

Mol. Cell. Biol. 5: 2051-2060(1985). Preferred cells for heterologous protein expression are those that can be readily and efficiently transfected.

Of particular interest is the use of this invention for the insertion of survival polynucleotides into the genome of murine embryonic stem (ES) cells. A preselection procedure based on the property of the survival polynucleotide can then be used to isolate specific gene-trapped ES cells that are either induced or repressed by a given modulator *in vitro* before generating the mice. Mouse strains with specific gene mutations due to insertion of survival polynucleotides and hence the interruption of the tagged genes in the host can be easily derived from a gene trap library constructed using embryonic stem cells, as mice can be bred to homozygosity to identify possible phenotypic changes caused by the mutation of the interrupted gene. These strains will help determine the role of the gene products that are either induced or repressed by a modulator in mammalian physiology and hence the relevance of these gene products to human disease (Zambrowicz B.P., Friedrich G.A. "comprehensive mammalian genetics; history and future prospects of gene trapping in the mouse." *Int J. Dev. Biol.* 1998 42:1025-1036).

Cells used in the present invention can be from continuous cell lines obtained from, for example, mammalian tissues, organs, or fluids. Primary cell lines can be made continuous using known methods, such as fusing primary cells with a continuous cell line or expressing transforming proteins. Cells of the invention can be stored or used with methods of the invention as isolated, clonal populations in plates. Preferably, cells are stored or used in plates with 96, 384, 1536 or 3456 wells per plate. A single cell or a plurality of cells can be placed in such wells. Such isolated clonal populations will typically have 1,000. 10,000, or 100,000 or more cells representative of substantially equivalent numbers of independent integrations sites. Such panels can be used in profiling, pathway

identification, modulator identification, modulator characterization, and other methods of the invention.

Prior to being transfected with a trapping vector of the present invention, cells can be transfected with an exogenous gene capable of expressing an exogenous protein, such as a receptor (e.g., GPCR) or gene associated with the pathology of an etiological agent, such as a virus, bacteria, or parasite. Cells that express such exogenous proteins can then be transfected with a trapping vector to form a library of clones that can be screened using the present invention.

Targets

Proteins of interest that can be expressed in the cells of the invention include,: hormone receptors (e.g. mineralcorticosteroid, gluococorticoid, and thyroid hormone receptors); intracellular receptors (e.g., orphans, retinoids, vitamin D3 and vitamin A receptors); signaling molecules (e.g., kinases, transcription factors, or molecules such signal transducers and activators of transcription) (*Science* Vol.264, 1994, p.1415-1421; *Mol. Cell Biol,* Vol.16, 1996, p.369-375); receptors of the cytokine superfamily (e.g. erthyropoietin, growth hormone, interferons, and interleukins (other than IL-B) and colony-stimulating factors); G-protein coupled receptors, see US patent 5,436,128 (e.g., for hormones, calcitonin, epinephrine, gastrin, and pancrine or autocrine mediators, such as stomatostatin or prostaglandins) and neurotransmitter receptors. (norepinephrine, dopamine, serotonin or acetylcholine); tyrosine kinase receptors (such as insulin growth factor, nerve growth factor (US patent 5,436,128). Examples of the use of such proteins is further described herein.

Any target, such as an intracellular or extracellular receptor involved in a signal transduction pathway, such as the leptin or GPCR pathways, can be used

with the present invention. Furthermore, the genes activated or repressed by a target can be isolated, identified, and modulators of that gene identified using the present invention. For example, the present invention can indentify a G-protein coupled receptor (GPCR) pathway, determine its function, isolate the genes modulated by the GPCR, and identify modulators of such GPCR modulated proteins.

In one aspect of the present invention, cells can be transformed to express an exogenous receptor, such as GPCR. Such a transduced cell line can than be further transduced with a trapping vector to make a library of clones that can be used to identify cells that report modulation of the exogenous receptor. Preferably, the host cell line would not appreciably express the exogenous receptor.

Based on the unique structure of GPCRs, which have seven hydrophobic, presumably trans-membrane, domains (see, Watson and Arkinstall. The G-Protein Linked Receptor Facts Book, Academic Press, New York (1994)) orphan GPCRs (GPCRs having no known function) can be identified by searching sequence databases, such as those provided by the National Library of Medicine (Bethesda, MD). for similar motifs and homologies. This same strategy can, of course, be used for any target, especially when a paradigm sequence or motif has been determined.

Drug Discovery for Viruses and Other Pathogens

The function of genes from viruses or other pathogens that effect the expression of genes in cells, such as mammalian cells, can be determined using the present invention. Furthermore, chemicals that modulate these genes can be identified using the methods of the present invention. For example, many

transforming viruses, after infecting a cell, have the effect of up-regulating genes involved in cell proliferation, which allows the virus-infected cells to produce additional viruses, which can infect additional cells. These transforming viruses can act by stimulating a receptor from the target cell. One example of the mechanism is the Friend Erythroleukemia virus. This virus uses the erythropoetin receptor for entry into the cells. When the virus is bound to the receptor, a pathway is activated that causes an over-proliferation of red blood cells. If the activation of the erythropoetin receptor is inhibited, a decrease in the accumulation of red blood cells would result which can prevent or reduce the severity of the leukemia. The development of an assay that reports the activation of mammalian target genes allows the identification of modulators of other viral or pathogenic dependent pathways. These modulators can be used as therapeutic agents.

A general procedure for establishing this assay uses the virus or an isolated viral protein as the stimulus for modulating a pathway. First, a gene-trapping library is made using a cell line that can be infected by the virus or activated by the viral protein. The virus is added to these cells, and clones are isolated that responded specifically to the viral infection by the expression of a reporter gene.

As an example, the GP120 portion of HIV protein is known to have mitogenic effect on cells exposed to GP-120, which indicates that downstream signaling pathways are being activated that can be associated with the cytotoxicity of the virus and allow its proliferation. Cell clones can be isolated that are induced by this activation which can be used to screen for modulators of this cytotoxic or proliferative effect. Other viral proteins, such as NEF from HIV, can be used. Chemicals that inhibit this effect can have useful therapeutic value to treat viral infection or toxicity.

This approach can be applied to any cellular pathogen that has an effect on target cells, such as cytotoxicity, cell proliferation, inflammation or other responses. Other etiological targets include other viruses, such as retroviruses, adenovirus, papiltomavirus, herpesviruses, cytomegalovirus, adeno-associated viruses, hepatitis viruses, and any other virus. In addition to viruses, any other pathogen, such as parasites, bacteria, and viroids, can be used in the present invention. Particular viral targets include, but are not limited to, NEF, Hepatitis X protein, and other viral proteins, such as those that can be encoded or carried by a virus. In addition, two or more viral components can be added to identify coviral pathogensis components. This is a particularly valuable tool for identifying pathways modulated by two or more viruses concurrently, or over time as in slow activating viral conditions. For example, cotransfection with HIV and CMV may be used. Viral targets or components do not include oncogenes or proto-oncogenes found in uninfected genomes, and gene products thereof.

Screening Test Chemicals Using Portions Of The Genome

Cells comprising survival polynucleotides integrated in the genome can be contacted with test chemicals or modulators of a biological process and screened for survival. Usually, the test chemical being screened will have at least one defined target, usually a protein. The test chemical is normally applied to the cells to achieve an appropriate concentration in the medium bathing the cells. Typically, screens are conducted at concentrations 100 μM or less, preferably 10 μM or less and preferably 1 μM or less for confirmatory screens. As described more fully herein, cells can be subjected to multiple rounds of screening and selection using the same chemical in each round to ensure the identification of clones with the desired response to a chemical or with different chemicals to characterize which chemicals produce a response (either survival or death) of the

cells. Such methods can be applied to any chemical that alters the function of any of the proteins mentioned herein or known in the art.

Chemicals and physiological processes without a defined target, however, can also be used and screened with the cells of the invention. For example, once a clone is identified as containing an active genomic polynucleotide that is activated by a particular cellular signal (including extracellular signals), for instance by a neurotransmitter, that same clone can be screened with chemicals lacking a defined target to determine if activation by the neurotransmitter is blocked or enhanced by the chemical. This is a particularly useful method for finding therapeutic targets downstream of receptor activation (in this case a neurotransmitter). Such methods can be applied to any chemical that alters the function of any the proteins mentioned herein or known in the art. This type of"targetless" assay is particular useful as a screening tool for the medical conditions and pathways described herein.

The methods and compositions described herein offer a number of advantages over the prior art. For instance, screening of mammalian based gene integration libraries is limited by the use of existing reporter systems. Many enzymatic reporter genes, such as luciferase, cannot be used to assay single intact living cells (for example by FACS) because the assay requires cell lysis to determine reporter gene activity.

Methods for Rapidly Identifying Modulators of Genomic Polynucleotides

The invention provides for a method of identifying proteins or chemicals that directly or indirectly modulate a genomic polynucleotide. Generally, the method comprises inserting a survival polynucleotide expression construct into an

eukaryotic genome, usually non-yeast, contained in at least one living cell, contacting the cell with a concentration of a modulator, and placing the cell under survival conditions. Preferably, the survival expression construct comprises a survival polynucleotide, a splice acceptor and an IRES element. The method can also include determining the coding nucleic acid sequence of a polynucleotide operably linked to the survival expression construct using techniques known in the art, such as RACE.

Modulator Identification

Modulators described herein can be used in this system to test for cell survival or death in successfully integrated clones. Such cells can optionally include specific proteins of interest as discussed herein. For example, the cell can include a protein or receptor that is known to bind the modulator (e.g., a nuclear receptor or receptor having a transmembrane domain heterologously or homologously expressed by the cell). A second modulator can be added either simultaneously or sequentially to the cell or cells and cell survival can be measured before, during or after such additions. Cells can be separated on the basis of their response to the modulator (e.g. responsive or non-responsive) and can be characterized with a number of different modulators to create a profile of cell activation or inhibition.

Cell survival will often be measured in relation to a reference sample, often a control. For example, cell survival is measured in the presence of the modulator and compared to the cell survival in the absence of the modulator or possibly a second modulator. Alternatively, cell survival is measured in a cell expressing a protein of interest and in a cell not expressing the protein of interest (usually the same cell type). For instance, a modulator may be known to bind to a receptor expressed by the cell and the survival of the cell is increased in the presence of the

modulator compared to the survival of a corresponding cell in the presence of the modulator, wherein the corresponding cell does not express the receptor.

Pathway Identification and Modulators

When a survival gene of the invention integrates into the genome of a host cell such that the survival gene is expressed under a variety of circumstances, these clones can be used for drug discovery and functional genomics. These clones report the modulation of the survival gene in response to a variety of stimuli, such as hormones and other physiological signals. These stimuli can be involved in a variety of known or unknown pathways that are modulated by known or unknown modulators or targets. Thus, these clones can be used as a tool to discover chemicals that modulate a particular pathway or to determine a cellular pathway.

These pathways are quite varied, and fall into general classes, which have specific species, which can be modulated by known or unknown modulators or agonists or antagonists thereof.

Extracellular signals from modulators regulate gene expression by triggering signal transduction cascades that result in the modulation of transcription factor activity. This is most commonly achieved through phosphorylation by signal responsive protein kinases. Phosphorylation affects transcription factor activity at several distinct levels. It can modulate their intracellular localization by controlling the association with other proteins, have both negative and positive effects on their DNA-binding activity, and modulate the activity of their transcriptional activation domains. In addition to phosphorylation, protein-protein interactions also have an important role in mediating a crosstalk at the nuclear level between different signaling pathways. (Karin M., "Signal transduction from the cell surface to the nucleus through the phosphorylation of

transcription factors" *Curr. Opin. Cell Biol.* 1994:6:415-424). With the advance in the molecular understanding of disease processes, it has been appreciated that many diseases result from the malfunctions of signaling pathways. This recognition has led to intensive research and the development of therapies based on the interception of cellular signaling in diseased cells. For instance success has been achieved using a blocker of the farnesylation of Ras as a tumor inhibitor, a JAK-2 blocker as an efficient inhibitor of recurrent pre-B cell acute lymphoblastic leukemia, and a platelet-derived growth factor receptor kinase as a blocker of restenosis (Levitzki A "Targeting signal transduction for disease therapy" Curr Opin Cell Biol. 1996: 8:239-244).

In one embodiment, the invention provides for a genomic assay system to identify downstream transcriptional targets for signaling pathways. This method requires the target of interest to activate gene expression upon addition of a chemical or expression of the target protein. A cell line that is the most similar to the tissue type where the target functions is preferred for generating a library of clones with different integration sites with survival polynucleotides. This cell line may be known to elicit a cellular response, such as differentiation upon addition of a particular modulator. If this type of cell line is available, it is preferred for screening, as it represents the native context of the target. If a cell line is not available that homologously expresses the target; a cell line can be generated by heterologously expressing the target in the most relevant cell line. For instance, if the target is normally expressed in the lymphoid cells, then a lymphoid cell line would be used to generate the library.

Once a pool of cells with the desired characteristics are isolated they can be expanded and their corresponding genes cloned and characterized. Targets that could be used in this assay system include receptors, kinases, protein/protein interactions or transcription factors and other proteins of interest discussed herein.

In another embodiment, the invention provides for a method of identifying developmentally or tissue specific expressed genes. Survival polynucleotides can be inserted, usually randomly, into any precursor cell such as an embryonic or hematopoetic stem cell to create a library of clones. Constitutively expressing clones can be collected by placing the cells under survival conditions. The library of clones can then be stimulated or allowed to differentiate, and induced or repressed clones isolated. Cell surface markers in conjunction with fluorescent tagged antibodies or other detector molecules could be used to monitor the expression of reference genes simultaneously. Additionally, by stimulating and sorting stem cells at various developmental stages, it is possible to rapidly identify genes responsible for maturation and differentiation of particular tissues.

Such methods can be used for identifying cell populations that have stem cells properties, as well as providing an intracellular reporter that allows isolation and screening of such a population of cells.

The present invention can yield cell lines for screening a variety of targets whose downstream signaling elements are already known or postulated. These screening cell lines can be used to either screen for modulators of transfected targets or as readouts for expression cloning or functional analysis of uncharacterized targets. Screening cell lines can be made for any pathway or any modulator.

Orphan protein signaling pathway identification and orphan protein modulators

In another embodiment, the invention provides for a method of identifying modulators of orphan proteins or genomic polynucleotides that are directly or indirectly modulated by an orphan protein. Human disease genes are often

identified and found to show little or no sequence homology to functionally characterized genes. Such genes are often of unknown function and thus encode an "orphan protein." Usually such orphan proteins share less than 25% amino acid sequence homology with other known proteins or are not considered part of a gene family. With such molecules there is usually no therapeutic starting point. By using libraries of the herein described clones, one can extract functional information about these novel genes.

Orphan proteins can be expressed, preferably overexpressed, in living mammalian cells. By inducing over-expression of the orphan gene and monitoring the effect on specific clones one may identify genes that are transcriptionally regulated by the orphan protein. By identifying genes whose expression is influenced by the novel disease gene or other orphan protein one may predict the physiological bases of the disease or function of the orphan molecule. Insights gained using this method can lead to identification of a valid therapeutic target for disease intervention.

Modulator Identification using Genomic Polynucleotides Activated by Cellular Signals

In another embodiment, the invention provides for a method of screening a defined target or modulator using genomic polynucleotides identified with the methods described herein. The gene identification methods described herein can also be used in conjunction with a screening system for any target that functions (either naturally or artificially) through transcriptional regulation.

In many instances a receptor and its ligand are known but not the downstream biological processes required for signaling. For example, a cytokine receptor and cytokine may be known but the downstream signaling mechanism is

not. A library of clones generated from a cell line that expresses the cytokine receptor can be screened to identify clones showing changes in gene expression when stimulated by the cytokine. The induced genes could be characterized to describe the signaling pathway. Using the methods of the invention, gene characterization is not required for screen development, as identification of a cell clone that specifically responds to the cytokine constitutes a usable secondary screen. Therefore, clones that show activation or deactivation upon the addition of the cytokine can be expanded and used to screen for agonists or antagonists of cytokine receptor. The advantage of this type of screening is that it does not require an initial understanding of the signaling pathway and is therefore uniquely capable of identifying leads for novel pathways.

In another embodiment, the invention provides for a method of functionally characterizing a target using a panel of clones having active genomic polynucleotides as identified herein. As large numbers of specifically responding cell lines containing active genomic polynucleotides identified with a particular biological process or modulator are generated, panels containing specific clones can be used for functional analysis of other potential cellular modulators. These panels of responding cell lines can be used to rapidly profile potential transcriptional regulators. Such panels, as well as containing clones with identified active genomic polynucleotides, which were generated by the invention panels, can include clones generated by more traditional methods. Clones can be generated that contain both the identified active genomic polynucleotide and specific response elements, such as SRE, CRE. NFAT, TRE, IRE, or reporters under the control of specific promoters. These panels would therefore allow the rapid analysis of potential effectors and their mechanisms of cellular activation.

In another embodiment, the invention provides for a method of test chemical profiling using a clone or panel of clones having identified active

polynucleotides. Test chemical characterization is similar to target characterization except that the cellular target(s) do not have to be known. This method will therefore allow the analysis of test chemical (e.g. lead drugs) effects on cellular function by defining genes effected by the drug or drug lead. Such a method can find useful applications in the area of drug discovery. The potential drug would be added to a library of genomic clones and clones which either were induced or repressed would be isolated, or identified. This method is analogous to target characterization except that the secondary drug target is unknown.

Once active genomic polynucleotides have been identified, they can be sequenced using various methods, including RACE (rapid amplification of cDNA ends). RACE is a procedure for the identification of unknown mRNA sequences that flank known mRNA sequences. 5' RACE is done by first preparing RNA from a cell line or tissue of interest. This total or polyA RNA is then used as a template for a reverse transcription reactions which can either be random primed or primed with a gene-specific primer. A poly nucleotide linker of known sequence is then attached to the 3' end of the newly transcribed cDNA by terminal transferase or RNA ligase. This cDNA is then used as the template for PCR using one primer within the reporter gene and the other primer corresponding to sequence which had been linked to the 3' end of the first stand cDNA. The present invention is particularly well suited for such techniques and does not require construction of additional clones or constructs once the genomic polynucleotide has been identified.

The present invention is also directed to chemical entities and information (e.g., modulators or chemicals or databases biological activities of chemicals or targets) generated or discovered by operation of the present invention, particularly chemicals and information generated using such systems.

Pharmacology and Toxicity of Candidate Modulators

The structure of a candidate modulator identified by the invention can be determined or confirmed by methods is known in the art, such as mass spectroscopy. For putative modulators stored for extended periods of time, the structure, activity. and potency of the putative modulator can be confirmed.

Depending on the system used to identify a candidate modulator, the candidate modulator will have putative pharmacological activity. For example, if the candidate modulator is found to inhibit T-cell proliferation (activation) *in vitro,* then the candidate modulator would have presumptive pharmacological properties as an immunosuppressant or anti-inflammatory (see, Suthanthiran et al., *Am. J. Kidney Disease*, 28:159-1 72 (1996)). Such nexuses are known in the art for several disease states, and more are expected to be discovered over time. Based on such nexuses, appropriate confirmatory in *vitro* and *in vivo* models of pharmacological activity, as well as toxicology, can be selected. The methods described herein can also be used to assess pharmacological selectivity and specificity, and toxicity.

Once identified, candidate modulators can be evaluated for toxicological effects using known methods (see, Lu, *Basic Toxicology, Fundamentals, Target Organs, and Risk Assessment*, Hemisphere Publishing Corp.. Washington (1985); U.S. Patent Nos: 5,196,313 to Culbreth (issued March 23, 1993) and U.S. Patent No.5,567,952 to Benet (issued October 22, 1996). For example, toxicology of a candidate modulator can be established by determining *in vitro* toxicity towards a cell line, such as a mammalian i.e. human, cell line. Candidate modulators can be treated with, for example, tissue extracts, such as preparations of liver, such as microsomal preparations, to determine increased or decreased toxicological properties of the chemical after being metabolized by a whole organism. The

results of these types of studies are often predictive of toxicological properties of chemicals in animals, such as mammals, including humans.

Alternatively, or in addition to these *in vitro* studies, the toxicological properties of a candidate modulator in an animal model, such as mice, rats, rabbits, or monkeys, can be determined using established methods (see, Lu, supra (1985); and Creasey, *Drug Disposition in Humans. The Basis of Clinical Pharmacology*, Oxford University Press, Oxford (1979)). Depending on the toxicity, target organ, tissue, locus, and presumptive mechanism of the candidate modulator, the skilled artisan would not be burdened to determine appropriate doses, LD_{50} values, routes of administration, and regimes that would be appropriate to determine the toxicological properties of the candidate modulator. In addition to animal models, human clinical trials can be performed following established procedures, such as those set forth by the United States Food and Drug Administration (USFDA) or equivalents of other governments. These toxicity studies provide the basis for determining the efficacy of a candidate modulator *in vivo*.

Efficacy of Candidate Modulators

Efficacy of a candidate modulator can be established using several art recognized methods, such as *in vitro* methods, animal models, or human clinical trials (see, Creasey, supra (1979)). Recognized *in vitro* models exist for several diseases or conditions. For example, the ability of a chemical to extend the life-span of HIV-infected cells *in vitro* is recognized as an acceptable model to identify chemicals expected to be efficacious to treat HIV infection or AIDS (see, Daluge et al., Antimicro. Agents Chemother. 41:1082- 1093 (1995)). Furthermore, the ability of a test chemical to prevent proliferation of T-cells *in vitro* has been established as an acceptable model to identify putative immunosuppressants (see, Suthanthiran et al., supra, (1996)). For nearly every class of therapeutic, disease,

or condition, an acceptable *in vitro* or animal model is available. Such models exist, for example, for gastro-intestinal disorders, cancers, cardiology, neurobiology, and imumunology. in addition, these *in vitro* methods can use tissue extracts, such as preparations of liver, such as microsomal preparations, to provide a reliable indication of the effects of metabolism on the candidate modulator. Similarly, acceptable animal models may be used to establish efficacy of chemicals to treat various diseases or conditions. For example, the rabbit knee is an accepted model for testing chemicals for efficacy in treating arthritis (see, Shaw and Lacy., *J. Bone Joint Surg. (Br)* 55:197-205(1973)). Hydrocortisone, which is approved for use in humans to treat arthritis, is efficacious in this model which confirms the validity of this model (see, McDonough, *Phys. Ther* 62:835-839 (1982)). When choosing an appropriate model to determine efficacy of a candidate modulator, the skilled artisan can be guided by the state of the art.

In addition to animal models, human clinical trials can be used to determine the efficacy of a candidate modulator in humans. The USFDA, or equivalent governmental agencies have established procedures for such studies.

Selectivity of Candidate Modulators

The *in vitro* and *in vivo* methods described above also establish the selectivity of a candidate modulator for a biological process. It is recognized that certain chemicals can modulate a wide variety of biological processes while others are selective for one or a few processes. Selective modulators may be preferable as chemotherapeutic agents because they have fewer side effects in the clinical setting. The selectivity of a candidate modulator can be assessed *in vitro* by using cell lines determined by the methods of this invention as described herein to exhibit particular signaling pathways. The data obtained from these studies can be extended to animal model studies and human clinical trials, to determine toxicity, efficacy, and selectivity of the candidate modulator.

The selectivity, specificity and toxicology, as well as the general pharmacology, of a test chemical can be often improved by generating additional test chemicals based on the structure/property relationships of the test chemical originally identified as having activity. Test chemicals identified as having activity can be modified to improve various properties, such as affinity, lifetime in the blood, toxicology, specificity and membrane permeability. Such refined test chemicals can be subjected to additional assays as described herein for activity analysis. Methods for generating and analyzing such chemicals are known in the art, such as U.S. patent 5,574,656 to Agrafiotis et al.

Compositions

The present invention also encompasses pharmaceutical compositions comprising a pharmaceutically effective amount of the chemical that has been identified by the methods of this invention as having modulating activity in a pharmaceutically acceptable carrier or diluent. Acceptable carriers or diluents for therapeutic use are well known in the pharmaceutical art, and are described, for example, in *Remington's Pharmaceutical Sciences*, 18th ed. Mack Publishing Co. (1990). Preservatives, stabilizers, dyes and even flavoring agents may be provided in the pharmaceutical composition. For example, sodium benzoate, sorbic acid and esters of p-hydroxybenzoic acid may be added as preservatives. In addition, antioxidants and suspending agents may be used.

The compositions of the present invention may be formulated and used as tablets capsules or elixirs for oral administration; suppositories for rectal administration; sterile solutions, suspensions for injectable administration; and the like. Injectables can be prepared in conventional forms either as liquid solutions or suspensions, solid forms suitable for solution or suspension in liquid prior to injection, or as emulsions. Suitable excipients are, for example, water, saline, dextrose, mannitol, lactose, lecithin, albumin, sodium glutamate, cysteine

hydrochloride, and the like. In addition, if desired, the injectable pharmaceutical compositions may contain minor amounts of nontoxic auxiliary substances, such as wetting agents, pH buffering agents, and the like, if desired, absorption enhancing preparations (e.g., liposomes), may be utilized.

The pharmaceutically effective amount or the candidate modulator required as a dose will depend on the route of administration, the age, weight and type of animal being treated, and the physical characteristics of the specific animal under consideration and the particular composition employed. The dose can be tailored to achieve a desired effect, but will depend on such factors as weight, diet, concurrent medication and other factors which those skilled in the medical arts will recognize (see e.g., Fingl et al in The Pharmacological Basis of Therapeutics, 1975). In practicing the methods of the invention, the pharmaceutical compositions can be used alone or in combination with one another, or in combination with other therapeutic or diagnostic agents. These products can be utilized *in vivo,* ordinarily in a mammal, preferably in a human, or *in vitro.* In employing them *in vivo,* the pharmaceutical composition can *be* administered to the mammal in a variety of ways, including parenterally, intravenously, subcutaneously, transdermally, transmucosally, intramuscularly, colonically, rectally, nasally or intraperitoneally, employing a variety of dosage forms. Such methods may also be applied to testing chemical activity *in vivo.*

The dosage for the products of the present invention can range broadly depending upon the desired affects and the therapeutic indication. Typically, dosages may be between about 10 ng/kg and 1 g/kg body weight, preferably between about 100 μg/kg and 10 mg/kg body weight. Administration is preferably oral on a daily basis.

For injection, the pharmaceutical compositions of the invention may be formulated in aqueous solutions, preferably in physiologically compatible buffers such as Hanks' solution, Ringer's solution, or physiological saline buffer. For such transmucosal administration, penetrants appropriate to the barrier to be permeated are used in the formulation. Such penetrants are generally known in the an. Use of pharmaceutically acceptable carriers to formulate the pharmaceutical compositions herein disclosed for the practice of the invention into dosages suitable for systemic administration is within the scope of the invention. With proper choice of carrier and suitable manufacturing practice, the compositions of the present invention, in particular, those formulated as solutions, may be administered parenterally, such as by intravenous injection. The pharmaceutical compositions can be formulated readily using pharmaceutically acceptable carriers well known in the an into dosages suitable for oral administration. Such carriers enable the chemicals of the invention to be formulated as tablets, pills, capsules, liquids, gels, syrups, slurries, suspensions and the like, for oral ingestion by a patient to be treated.

Agents intended to be administered intracellularly may be administered using techniques well known to those of ordinary skill in the art. For example, such agents maybe encapsulated into liposomes, then administered as described above. All molecules present in an aqueous solution at the time of liposome formation are incorporated into the aqueous interior. The liposomal contents are both protected from the external micro- environment and, because liposomes fuse with cell membranes, are efficiently delivered into the cell cytoplasm. Additionally, due to their hydrophobicity, small organic molecules may be directly administered intracellularly.

Pharmaceutical compositions suitable for use in the present invention include compositions wherein the active ingredients are contained in an effective

amount to achieve its intended purpose. Determination of the effective amount of a pharmaceutical composition is well within the capability of those skilled in the art, especially in light of the detailed disclosure provided herein. In addition to the active ingredients, these pharmaceutical compositions may contain suitable pharmaceutically acceptable carriers comprising excipients and auxiliaries which facilitate processing of the active chemicals into preparations which can be used pharmaceutically. The preparations formulated for oral administration may he in the form of tablets, dragees, capsules, or solutions. The pharmaceutical compositions of the present invention may be manufactured in a manner that is itself known, e.g., by means of conventional mixing, dissolving, granulating. dragee-making, emulsifying, encapsulating, entrapping, or lyophilizing processes. Pharmaceutical formulations for parenteral administration include aqueous solutions of the active chemicals in water-soluble form. Additionally, suspensions of the active chemicals may be prepared as appropriate oily injection suspensions. Suitable lipophilic solvents or vehicles include fatty oils such as sesame oil, or synthetic fatty acid esters, such as ethyl oleate or triglycerides, or liposomes. Aqueous injection suspensions may contain substances which increase the viscosity of the suspension, such as sodium carboxymethyl cellulose, sorbitol, or dextran. Optionally. the suspension may also contain suitable stabilizers or agents that increase the solubility of the chemicals to allow for the preparation of highly concentrated solutions.

Pharmaceutical compositions for oral use can be obtained by combining the active chemicals with solid excipient, optionally grinding a resulting mixture, and processing the mixture of granules. after adding suitable auxiliaries, if desired, to obtain tablets or dragee cores. Suitable excipients are, in particular, fillers such as sugars, including lactose, sucrose, mannitol, or sorbitol; cellulose preparations such as, for example, maize starch, wheat starch, rice stanch, potato starch, gelatin, gum tragacanth, methyl cellulose, hydroxyprnpylmethyl-cellulose sodium

carboxymethylcellulose, an/or polyvinylpyrro'idone (PVP). If desired, disintegrating agents may be added. such as the cross-linked polyvinyl pyrrolidone, agar, or alginic acid or a salt thereof such as sodium alginate. Dragee cores are provided with suitable coatings. For this purpose, concentrated sugar solutions may be used, which may optionally contain gum arabic, talc, polyvinyl pyrrolidone, carbopol gel, polyethylene glycol, and/or titanium dioxide, lacquer solutions, and suitable organic solvents or solvent mixtures. Dye-stuffs or pigments may be added to the tablets or dragee coatings for identification or to characterize different combinations of active chemical doses.

In order to further illustrate the present invention and advantages thereof, the following specific examples are given but are not meant to limit the scope of the claims in any way.

EXAMPLES

In the examples below, all temperatures are in degrees Celsius (unless otherwise indicated) and all percentages are weight percentages (also unless otherwise indicated).

In the examples below, the following abbreviations have the following meanings. If an abbreviation is not defined, it has the generally accepted meaning:

μM	=	micromolar
mM	=	millimolar
M	=	molar
μL	=	microliter
mL	=	milliliter
μg	=	microgram
mg	=	milligram

bp	=	base pair
ng	=	nanogram
PAGE	=	polyacrylamide gel electrophoresis
SDS	=	sodium dodecyl sulfide
PBS	=	phosphate buffered saline
IRES	=	internal ribosome entry bidning site
EGFP	=	enhanced green fluorescent protein
Kan^R	=	kanamycin resistance gene
Neo^R	=	neomycin resistance gene
$Puro^R$	=	puromycin resistance gene
HSV TK	=	Herpes Simplex virus thymidine kinase gene

General Methods:

Standard molecular biology methods known in the art and not specifically described were generally followed as in Sambrook et al., (1992); in Ausubel et al., (1989) "Current Protocols in Molecular Biology" John Wiley & Sons, Baltimore MA, and in Perbal (1988) "A Practical Guide to Molecular Cloning" John Wiley and Sons, New York. Polymerase chain reaction (PCR) was carried out generally as in PCR protocols: A Guide to Methods and Applications, Academic Press, San Diego CA. (1990)

Example 1: Construction of pTadpole

The pTadpole vector was constructed by inserting the IRES-$Puro^R$ portion of pIRESpuro into the mcs of pIRES2-EGFP. Figure 2A shows the restriction map of pIRESpuro (Clontech) and Figure 2B shows the restriction map of pIRES2-EGFP (Clontech).

pIRESpuro was digested with NsiI and BclI. The 1354 bp IRES-$Puro^R$ fragement was isolated.

pIRES2-EGFP was digested with PstI and BamHI and run through a gel. The 5286 bp vector was isolated from the gel.

The NsiI/BclI IRES-Puro^R fragment was ligated into the PstI/BamHI sites of pIRES2-EGFP to create pTadpole. Compatible cohesive ends are NsiI/PstI and BclI/BamHI. pTadpole constitutively expresses neomycin from the SV40 promoter and puromycin and EGFP from the CMV_{IE} promoter.

Example 2: Construction of pFrog-CMV and pFrog-PCV

pFrog-CMV was generated by replacing the CMV_{IE} promoter of pTadpole with the CMV_{IE}/EF-1α promoter and EF-1α intron 1 splice acceptor from pCE2

The plasmid pCE2 (Weeks et al., *DNA Cell Biol.* 16:281-289 1997) was derived from pREP7b (Leung et al., *Proc. Natl. Acad. Sci, USA* 92:4813-4817 1995) with the RSV promoter region replaced by the CMV enhancer and the elongation factor 1α (EF-1α) promoter and intron. The CMV enhancer came from a 380 bp XbaI- SphI fragment produced by PCR from pCEP4 (Invitrogen, San diego CA) using the primers 5'-GGCTCTAGAT ATTAATAGTA ATCAATTAC-3' and 5'-CCTCACGCAT GCACCATGGT AATAGC-3'. The EF-1α promoter and intron (Uetsuki et al., *J. Biol. Chem.* 264:5791-5798 1989) came from a 1200 bp SphI-Asp718 fragment produced by PCR from human genomic DNA using the primers 5'-GGTGCATGCG TGAGGCTCCG GTGC-3' and 5'-GTAGTTTTCA CGGTACCTGA AATGGAAG-3'. These 2 fragments were ligated into a XbaI/Asp718 digested vector derived from pREP7b to generate pCE2.

Specifically, pCE2 was digested with SnaBI and BamHI and run through a gel. A 1261 bp fragment containing the CMV_{IE}/EF-1α promoter;EF-1α intron1 splice acceptor fragment was isolated. pTadpole was digested with SnaBI and

BglII, run through a gel and a 6371 bp vector band was isolated. The SnaBI/BamHI CMV_{IE}/EF-1α promoter;EF-1α intron splice acceptor fragment was ligated into the SnaBI/BglII sites of pTadpole to create pFrog-CMV. A schematic map of pFrog-CMV is shown in Figure 3.

pFrog-PCV was constructed by removing the constitutive CMV_{IE}/EF-1α promoter from pFrog-CMV while leaving the majority of the EF-1α intron 1 splice acceptor site intact. pFrog-PCV can be used as a promoter trapping vector by linearizing the vector at BglII or AlwNI prior to transfection.

Specifically, pFrog-CMV was digested with AseI and BglII and a 6692 bp vector fragment was isolated. The overhanging ends from the restriction digest of the fragment were filled in using Klenow fragment. The now blunt ended vector was recircularized with T4 ligase to create pFrog-PCV. This recircularization regenerates the BglII site but the AseI site is destroyed. Figure 4A is a schematic map of the pFrog-PCV. Figure 4B shows the vector as it would appear when integrated into the eukaryotic genome. pFrog-PCV constitutively expresses neomycin from the SV40 promoter. However, expression of puromycin and EGFP is dependent upon endogenous promoter elements. Hence the vector can function as a promoter trapping vector.

Example 3: Expression of pFrog-CMV and pFrog-PCV in eukaryotic cells.

The vectors were introduced into eukaryotic cells to determine whether expression could be detected.

Specifically, 5 μg of pFrog-CMV, pFrog-PCV (linearlized with AlwNI (CMV) or BglII (PCV) were electroporated individually into 5 x 10^6 cells of ECV304 (American Type Culture Collection, Manassas VA) with a Cell-Porator™

(Life Technologies, Gaithersburg MD) using conditions described previously (Cachianes et al., *Biotechniques* 15:255-259 1993). One day later the 400 μg/ml of G418 was added to the cells. Four days later the concentration of G418 was increased to 600 μg/ml. This level of G418 resulted in cell death in ECV304 cells not having either plasmid by the seventh day. The concentration of G418 was increased to 800 μg/ml on the eighth day. A this time, individual pFrog-CMV and pFrog-PCV transfectants were apparent.

Single transfectants were picked and plated on duplicate plates. One plate was placed under 300 μg/ml of puromycin. The second plate was placed under 600 μg/ml of puromycin. For both vectors, 100 % survival was observed at 300 μg/ml of puromycin and 50% survival was observed at 600 μg/ml of puromycin. All of the surviving colonies were observed under the microscope for green fluorescent protein. All of the pFrog-CMV transformants were bright green. However, less than 5% of the pFrog-PCV transformants were green.

The results of this example established that all of the selectable markers in the vectors were able to select for certain integrated plasmids.

Example 4: Construction of pSOF-CMV, pSOF-PCV and pSOF-IL6

The following vectors provide both negative and positive selection. This series of vectors replaces the $Puro^R$ gene of pFrog with the tk::shble gene from pGT65hIFNα (Invivogen, San Diego CA).

Figure 5 is the nucleotide sequence of HSV1-thymiding kinase gene fused to the zeocin resistance gene. A 1583 bp section of the thymidine kinase:zeocin resistance fusion gene was obtained by PCR amplification of the pGT65HIFNα

plasmid DNA using a TK-shble-F/R primer set using conditions described by Boehringer Mannheim Expand Long template PCR system.

Forward primer for construction of promoter capture vectores 5' EspI (BsmBI) site allows fusion of TK-shble to the first IRES in pFrog-CMV and pFrog-PCV. When used in PCR of pGT65hIFNa with TK-shble-R the amplified product encodes the HSV-1 thymiding kinase fused to the zeocin resistance gene.

Forward Primer

5'-atgcatacaa ggagacgacc ttccATGTCG ACTACTAACC TTC-3'

Reverse Primer for construction of promoter capture vector(s) 3' BamHI/XbaI site allows fusion of TK-shble tothe first IRES in pFrog-CMV and pFrog-PCV.

Reverse Primer

5'-atgcatctag aggatccTCA GTCCTGCTCC TCGGCCACGA AG-3'.

The resulting 1583 bp PCR product was restriction digested with EspI (BsmBI) for the 5'-end and with XbaI for the 3' end.

pFrog-CMV was restriction digested with BsmBI and partial XbaI (the other XbaI site is methylated when the plasmid is is *E coli* DH10B cells). A 6897 bp vector band was isolated from the pFrog-CMV digestion. This fragment was ligated with the BsmBI/XbaI thymidine kinase:zeocin resistance fusion gene product fragment to create pSOF-CMV. Figure 6A is the schematic map of pSOF-CMV. Figure 6B shows the vector once it is integrated into the eukaryotic genome.

pSOF-CMV constitutively expresses neomycin from the SV40 promoter and constitutively expresses thymidine kinase:zeocin resistance and EGFP from the CMV_{IE}/EF-1α promoter. For use as a constitutive plasmid, pSOF-CMV can be linearized at AseI (or AlwNI) prior to transfection into eukaryotic cells.

pSOF-PCV is the result of replacing the Puro^R reporter gene of pFrog-PCV with the tk:sh ble reporter gene from pGT65hIFNα. pSOF-PCV can be linearized at BglII (or AlwNI) prior to transfection for use as a promoter trapping vector.

The PCR amplified 1583 bp thymidine kinase::zeocin resistance fusion gene product from pGT65hIFNα plasmid DNA was used. This fragment was digested with EspI (BsmBI) for the 5' end and with XbaI at the 3' end. pFrog-PCV was digested with BsmBI and partially digested with XbaI (the other XbaI site is methylated in *E. coli* DH10B cells). A 5959 bp vector band was isolated from the digestion of pFrog-PCV. This fragment was ligated with the BsmBI/XbaI sites of the thymidine kinase::zeocin resistance fusion gene product fragment to create pSOF-PCV.

Figure 7A is the schematic map of pSOF-PCV. Figure 7B shows the vector once it is integrated into the eukaryotic genome. pSOF-PCV constitutively expresses the neomycin resistance gene from the SV40 promoter. It shows bicistronic expression of the thymidine kinase::zeocin gene and EGFP if placed next to a promoter after transfection into a eukaryotic cell genome. Therefore, this plasmid can be used as a promoter trapping vector.

pSOF-IL6 is the result of adding the IL-1 inducible promoter from pIL6.AP just upstream of the reporter cassette in pSOF-PCV. pSOF-PCV can be linearized at the restriction site for AlwNI prior to transfection for use as an inducible reporter plasmid.

To construct pIL6.AP the fragment encoding the promoter for IL-6 was amplified from human genomic DNA (Promega Madison WI) by PCR using the primers 5'-GGGCCTCTAG ACTGTTAATC TGGTC-3' and 5'-CAGCTGGTAC

CGGTGGCTCG AGGGGCAGAAT G-3'. The resultant PCR product was then digested with XbaI and Acc65 I to generate a 440 bp fragment. The 440 bp fragment was then inserted into the pREP7b.AP vector cleaved with XbaI and Acc65I to generate pIL6.AP (Leung et al., (1995) PNAS 921:4813-4817)

To construct pSOF-IL6, the plasmid pIL6.AP was digested with XbaI and XhoI. The resulting fragments were blunt-ended with T4 DNA polymerase. The 426 bp IL6 promoter fragment was isolated.

The plasmid pSOF-PCV was digested with BglII. The digested product was made blunt ended with T4 DNA polymerase and a 7526 vector band was isolated by agarose electrophoresis. The blunt-ended IL6 fragment was ligated into the blunt-ended BglII site of pSOF-PCV to create pSOF-IL6.

Figure 8A is a schematic map of the plasmid pSOF-IL6. Figure 8B shows the configuration of the plasmid after it has been linearized and transfected into eukaryotic genomic DNA. pSOF-IL6 has constitutive neomycin expression from the SV40 promoter and inducible bicistronic expression of the thymidine kinase:zeocin fusion gene and EGFP from the inducible IL6 promoter. pSOF-IL6 can be linearized at AlwNI prior to transfection for use as a control vector containing an inducible promoter.

Example 5: Generation of an ECV304 transfectant libraries using pSOF-PCV or pSOF-IL6

The plasmid pSOF-PCV was used to transfect ECV 304 cells to generate a promoter trapping library.

The plasmid pSOF-PCV was linearized by digestion of the DNA with the restriction enzyme BglII. Approximately 0.6 μg of linearized pSOF-PCV DNA was transfected into 5 x 10^6 cells by electroporation with a Cell Porator™ (Life Technologies, Gaithersburg MD) using conditions described previously. The transfected ECV304 cells were allowed to recover overnight and then were exposed to 800 μg/ml of G418. On the eighth day, single transfectants were identified and placed in fresh medium without G418. Similarly ECV304 cells were transfected with pSOF-IL6.

The isolated pSOF-PCV and pSOF-IL6 transfectants were induced by the addition of 0.5 ng/ml IL-1β. The cells were kept in the IL-1β solution for 1.5 hours and then the cellular RNA was harvested. Northern blots were run of the cellular RNA. Figure 9A shows the procedure for validating the expression of the tk:sh ble gene in the pSOF-PCV transfectants. The northern blots were probed with radiolabelled tk:sh ble.

Figure 9B illustrates the Northern blot after is was probed with radiolabelled tk:sh ble. SOF-IL6 refers to transfectants incorporating pSOF-IL6. SOF-IL1 refers to transfectants incorporating pSOF-PCV, which were later found to have inducible IL-1β promoters. The lower band (right side) in the upper gel is artifactual ribosomal bands. By comparison with the upper band (left side), the lower band for PCV-IL1 panels (right side) appears to be smaller than the upper band on the left panel. This is possibly because of a deletion of a small part of the vector during transfection (or alternatively could be due to a strong structure effect of IRES that affects the mobility of the band). However, the tk-sh ble gene is a full length as determined by PCR analysis. G3PDH (glyceraldehyde-3-phosphate dehydrogenase) is an internal control to measure the mount of RNA loaded on the gel.

A pSOF-IL6 transfectant SOF-IL6.9 was picked as was a pSOF-PCV transfectant PCV-il1.2w. These transfectants were grown and then induced by the addition of 0.5 ng/ml of IL-1β. The level of RNA transcribed after indcution was measured. Figure 9C is a graph showing the increase in the level of transcription observed between the different transfectants after induction. (-) before induction (+) after induction.

Example 6: Time Course Analysis of Induction of transfectants with IL-1β

The IL-1 inducible control transfectant, SOF-IL6.9 was grown and induced by the addition of 0.5 ng/ml of IL-1β. Cells were harvested at 0, 0.5, 1, 2, 4, 5, and 20 hours post treatment and total RNA was isolated (Figure 10A). the RNAs (5 μg each time point) were resolved on denaturing agarose gels, transferred to nylon membranes and probed with radiolabeled DNA probes to tk::shble or G3PDH (Figure 10B). The levels of tk::shble and G3PDH expression were quantitated by phosphorimager and the results (normalized to background) are shown (Figure 10C).

The IL-1β-inducible transfectant PCV-IL1.2z was grown and similarly induced by the addition of 0.5 ng.ml of IL-1β for 0, 0.5, 1, 2, 4, 6 and 20 hours. Following treatment, cells were harvested and total RNA was isolated (Figure 11A). The RNAs (5 μg each time point) were resolved on denaturing agarose gels, transferred to nylon membranes and probed with radiolabeled DNA probes to tk::shble or G3PDH (Figure 11B). The levels of tk::shble and G3PDH expression were quantitated by phosphorimager and the results (normalized to background) are shown in Figure 11C.

Example 7: Construction of pDOF-CMV, pDOF-PCV and pDOF-IL6

The following vectors provide both negative and positive selection. This series of vectors replaces the sh ble protion of the tk::sh ble fusion gene in pSOF-CMV withthe bsdS portion of pUB6V5HB. (Invitrogen Carlsbad CA). The shble gene product imparts bleomycin or zeocin resistance in cells by binding directly to zeocin. Hence one molecule of shble is required to neutralize one molecule of zeocin introduced. On the contrary bsd imparts blasticidinS resistance in cells by hydrolyzing blasticidinS. Hence one molecule of bsd gene product is capable of neutralizing multiple molecules of blasticidinS due to bsd's enzymatic activity. Replacing bsd with shble should increase the sensitivity of selecting the surviving cells.

A 416 bp section of the blasticidin S resistance gene was obtained by PCR amplification of the pUB6V5HB plasmid DNA using a BSD-F/R primer set.

Forward Primer

Forward primer for construction of promoter capture vector(s). 5' BclII site allows fusion of TK in pSOF series vectors to replace the sh ble (zeocin resistance) gene with BSD. when used in pCR of pUB6/V5-HisB with BSD-R1, the amplified product encodes the blasticidin S resistance gene of Aaspergillis.
5'-atgcattgat cagcCCTTTG TCTCAAGAAG AATC-3'

Reverse Primer

Reverse primer for construction of promoter capture vector(s) 3' XbaI site alows fusion of TK in pSOF series vectors to replace the sh ble (zeocin resistance) gene with BSD. when used in pCR of pUB6/V5-HisB with BSD-F1 the amplified product encodes the basticidin S resistance gene of Aspergillis.
5' atgcattcta gaTTAGCCCT CCCACACATA ACCAG-3'

The resulting 416 bp PCR product was restriction digested with BclI for the 5'-end and with XbaI for the 3' end.

pSOF-CMV was restriction digested with BclI and partial XbaI. A 8060 bp vector band was isolated from the pSOF-CMV digestion. This fragment was ligated with the BclI/XbaI bsdS gene product fragment to create pDOF-CMV. Figure 12A is the restriction map of pDOF-CMV. Figure 12B shows the vector in transfectants after integration into the eukaryotic genome.

pDOF-CMV constitutively expresses neomycin from the SV40 promoter and constitutively expresses thymidine kinase:blasticidin and EGFP from the CMV_{IE}/EF-1α promoter. For use as a constitutive plasmid, pDOF-CMV can be linearized at AseI (or AlwNI) prior to transfection into eukaryotic cells.

pDOF-PCV is the result of replacing the sh ble portion of the tk::sh ble fusion in pSOF-PMV with the bsdS portion of pUB6V5HB. pDOF-PCV can be linearized at BglII (or AlwNI) prior to transfection for use as a promoter trapping vector.

The PCR amplified 416 bp blasticidin S resistance gene product from pUB6V5HB plasmid DNA was used. This fragment was digested with BclI for the 5' end and with XbaI at the 3' end. pSOF-PCV was digested with BclI and partially digested with XbaI (the other XbaI site is methylated in *e. coli* DH10B cells). A 7122 bp vector band was isolated from the digestion of pSOF-PCV. This fragment was ligated with the BclI/XbaI sites of the blasticidinS resistance gene product fragment to create pDOF-PCV.

Figure 13A is a schematic map of pDOF-PCV. Figure 13B shows the vector once it is integrated into the eukaryotic genome. pDOF-PCV constitutively

expresses the neomycin resistance gene from the SV40 promoter. It shows bicistronic expression of the thymidine kinase::blasticidin gene and EGFP if placed next to an inducible promoter after transfection into a eukaryotic cell genome. Therefore, this plasmid can be used as a promoter trapping vector.

pDOF-IL6 is the result of replacing the sh ble portion of the tk::sh ble fusion in pSOF-IL6 with the bsdS portion of pUB6VHB. pDOF-IL6 can be linearized at the restriction site for AlwNI prior to transfection for use as an inducible reporter plasmid.

The PCR amplified 416 bp blasticidin S resistance gene product from pUB6V5HB plasmid DNA was used. This fragment was digested with BclI for the 5' end and with XbaI at the 3' end. pSOF-IL6 was digested with BclI and partially digested with XbaI. A 7566 bp vector band was isolated from the digestion of pSOF-IL6. This fragment was ligated with the BclI/XbaI sites of the blasticidinS resistance gene product fragment to create pDOF-IL6.

Figure 14A is a schematic map of the plasmid pDOF-IL6. Figure 14B shows the configuration of the plasmid after it has been linearized and transfected into eukaryotic genomic DNA. pDOF-IL6 has constitutive neomycin expression from the SV40 promoter and inducible bicistronic expression of the thymidine kinase:blasticidin gene and EGFP from the IL6 promoter, thus is an inducible control vector.

Example 8: Construction of pICOF-CMV, pICOF-PCV and pICOF-IL6

The following vectors provide both negative and positive selection. This series of vectors replaces the EGFP reporter gene of pDOF-PCV in pSOF-PCV with the secreted human placental alkaline phosphatase reporter (SEAP) gene from

pIL6.AP. The pIL6.AP was derived from the plasmid pREP7b-AP (Leung et al., 1995 Proc. Natl. Acad. Sci. 92:4813-4817.

The plasmid pIL6.AP was digested with ACC65I and HpaI. The digested products were made blunt-ended with T4 DNA polymerase. A 1550 bp DNA fragment having the SEAP gene was isolated by agarose gel electrophoresis.

pDOF-PCV was digested with BstXI and HpaI. The digested product was made blunt-ended with T4 DNA polymerase. The 6678 bp DNA fragment corresponding to the vector band was isolated. The blunt-ended 1550 bp DNA fragment having the SEAP gene was blunt ligated into the BstXI/HpaI sites of pDOF-PCV to create pICOF-PCV.

Figure 15A is a schematic map of pICOF-PCV. Figure 15B shows the vector in transfectants after integration into the eukaryotic genome.

pICOF-PCV constitutively expresses neomycin from the SV40 promoter. The plasmid inducibly expresses thymidine kinase:blasticidin and SEAP. For use as a promoter trapping plasmid, pICOF-PCV can be linearized with BglII prior to transfection into eukaryotic cells. Alternatively, the plasmid could be linearized at other sites, for example the DRAI site on the plasmid. Therefore, this plasmid can be used as a promoter trapping vector by the methods set forth in the examples above.

pICOF-CMV is the result of adding the CMV_{IE}/EF-1α promoter from pSOF-CMV just upstream of the reporter cassette in pICOF-PCV. pICOF-CMV can be linearized at the AseI site or at the DraI site prior to transfection for use as a constitutive expression vector.

pSOF-IL6 was digested with FspI and EcoRI. A 3495 bp CMV_{ie}/EF-1α promoter containing fragment was isolated from the digestion of pSOF-IL6. pICOF-PCV was digested with FspI and EcoRI. A 5667 bp DNA fragment having the vector was isolated. The 3495 bp CMV_{ie}/EF-1α promoter fragment was ligated with the vector band fragment to create pICOF-CMV.

Figure 16A is a schematic map of pICOF-CMV. Figure 16B shows the vector once it is integrated into the eukaryotic genome. pICOF-CMV constitutively expresses the neomycin resistance gene from the SV40 promoter and constitutively expresses the thymidine kinase::blasticidin gene and SEAP from the CMV_{ie}/EF-1α promoter after transfection into a eukaryotic cell genome.

pICOF-IL6 is the result of adding the IL-1 inducible promoter from pIL6.AP just upstream of the reporter cassette in pICOF-PCV. pICOF-IL6 can be linearized at the restriction site Eco47III or at the DraI site prior to transfection for use as an inducible reporter plasmid.

The plasmid pICOF-IL6 was constructed by digesting pSOF-IL6 with FspI and EcoRI. A 2991 bp DNA fragment having the IL-6 promoter was isolated. The pICOF-PCV plasmid was also digested with FspI and EcoRI and a 5667 bp DNA fragment having the vector was isolated. The two isolated fragments were ligated to create pICOF0IL6.

Figure 17A is a schematic map of the plasmid pICOF-IL6. Figure 17B shows the configuration of the plasmid after it has been linearized and transfected into eukaryotic genomic DNA. pICOF-IL6 has constitutive neomycin expression from the SV40 promoter and inducible bicistronic expression of the thymidine kinase:blasticidin gene and SEAP from the IL6 promoter, thus is an IL-1 inducible control vector.

Claims:

1. A method for identifying modulators that directly or indirectly modulate expression of a genomic polynucleotide comprising:

providing a nucleic acid sequence comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site integrated into a genomic polynucleotide in a eukaryotic genome contained in at least one living cell which survival polynucleotide is transcriptionally incompetent,

contacting said cell with a predetermined concentration of a modulator, and

placing the cell under survival conditions and identifying those cells which survive.

2. The method of Claim 1 wherein domain 1 of the survival polynucleotide is selected from the group consisting of the zeocin gene, hygromycin gene, neomycin gene, blasticidin S, puromycin gene.

3. The method of Claim 1 wherein domain 2 of the survival polynucleotide is selected from the group consisting of the thymidine kinase gene and the cytidine deaminase gene.

4. The method of Claim 1 wherein said living cell is a mammalian cell.

5. The method of Claim 1 wherein said modulator is a peptide.

6. The method of Claim 1 wherein said modulator is an agonist.

7. The method of Claim 1, wherein said modulator is an antagonist.

8. The method of Claim 1 wherein the survival conditions kill those cells which do not transcribe the survival polynucleotide.

9. The method of Claim 1 wherein the survival conditions kill those cells which do transcribe the survival polynucleotide.

10. A method for identifying modulators, comprising:

(a) providing a nucleic acid sequence comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site and a known inducible promoter, which sequence is integrated into a eukaryotic genome contained in at least one living cell,

(b) contacting said cell with a predetermined concentration of a test chemical, and

(c) placing the cell under survival conditions and identifying those cells which survive.

11. The method of Claim 10 further comprising:

(d) providing the nucleic acid sequence as set forth in step (a) which sequence is integrated into a eukaryotic genome contained in at least one living cell,

(e) contacting said cell with a predetermined concentration of a known modulator,

(f) placing the cell under survival conditions and identifying those cells which survive, and

(g) determining whether the percentage of cells that survive step (c) is comparable to the percentage of cells that survive step (f).

12. The method of Claim 10 wherein domain 1 of the survival polynucleotide is selected from the group consisting of the zeocin gene, hygromycin gene, neomycin gene, blasticidin S, puromycin gene.

13. The method of Claim 10 wherein domain 2 of the survival polynucleotide is selected from the group consisting of the thymidine kinase gene and the cytidine deaminase gene.

14. The method of Claim 10 wherein said living cell is a mammalian cell.

15. The method of Claim 10 wherein said modulator is a peptide.

16. The method of Claim 10 wherein said modulator is an agonist.

17. The method of Claim 10, wherein said modulator is an antagonist.

18. The method of Claim 10 wherein the survival conditions kill those cells which do not transcribe the survival polynucleotide.

19. The method of Claim 10 wherein the survival conditions kill those cells which do transcribe the survival polynucleotide.

20. A method for identifying intracellular pathways, comprising:

providing a plurality of eukaryotic cells, wherein the eukaryotic genome of each cell comprises a nucleic acid sequence comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site and a known inducible promoter, wherein said plurality of cells has a plurality of integration sites where said nucleic acid sequence has integrated,

contacting said plurality of eukaryotic cells with a modulator of interest,

placing the plurality of cells under survival conditions and identifying those cells which survive,

wherein survival of said cells indicates participation of said integration site in the intracellular pathway.

21. The method of Claim 20, wherein said eukaryotic cell is a mammalian cell.

22. The method of Claim 20 wherein domain 1 of the survival polynucleotide is selected from the group consisting of the zeocin gene, hygromycin gene, neomycin gene, blasticidin S, puromycin gene.

23. The method of Claim 20 wherein domain 2 of the survival polynucleotide is selected from the group consisting of the thymidine kinase gene and the cytidine deaminase gene.

24. The method of Claim 20 wherein said modulator is a peptide.

25. The method of Claim 20 wherein said modulator is an agonist.

26. The method of Claim 20, wherein said modulator is an antagonist.

27. The method of Claim 20 wherein the survival conditions kill those cells which do not transcribe the survival polynucleotide.

28. The method of Claim 20 wherein the survival conditions kill those cells which do transcribe the survival polynucleotide.

29. A method for identifying a promoter region capable of being modulated by a modulator, comprising:

providing a plurality of eukaryotic cells, wherein the eukaryotic genome of each cell comprises a nucleic acid sequence comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site, wherein said plurality of cells has a plurality of integration sites where said nucleic acid sequence has integrated,

contacting said plurality of eukaryotic cells with a modulator of interest,

placing the plurality of cells under survival conditions and identifying those cells which survive, and

isolating the promoter region at the integration site operably linked to the survival polynucleotide in the surviving cells.

30. A method for identifying an enhancer region capable of being modulated by a modulator, comprising:

providing a plurality of eukaryotic cells, wherein the eukaryotic genome of each cell comprises a nucleic acid sequence comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a known weak promoter region requiring an enhancer, a splice acceptor site and an internal ribosome entry binding site, wherein said plurality of cells has a plurality of integration sites where said nucleic acid sequence has integrated,

contacting said plurality of eukaryotic cells with a modulator of interest,

placing the plurality of cells under survival conditions and identifying those cells which survive, and

isolating the enhancer region operably linked to the survival polynucleotide in the surviving cells.

31. An ES cell comprising a nucleic acid sequence integrated into the genome of the cell comprising a survival polynucleotide comprising a domain 1 and a

domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site.

32. A plurality of ES cells each comprising a nucleic acid sequence integrated into the genome of the cell comprising a survival polynucleotide comprising a domain 1 and a domain 2 operably linked to a splice acceptor site and an internal ribosome entry binding site wherein said plurality of cells has a plurality of integration sites where said nucleic acid sequence has integrated.

FIG. 1A

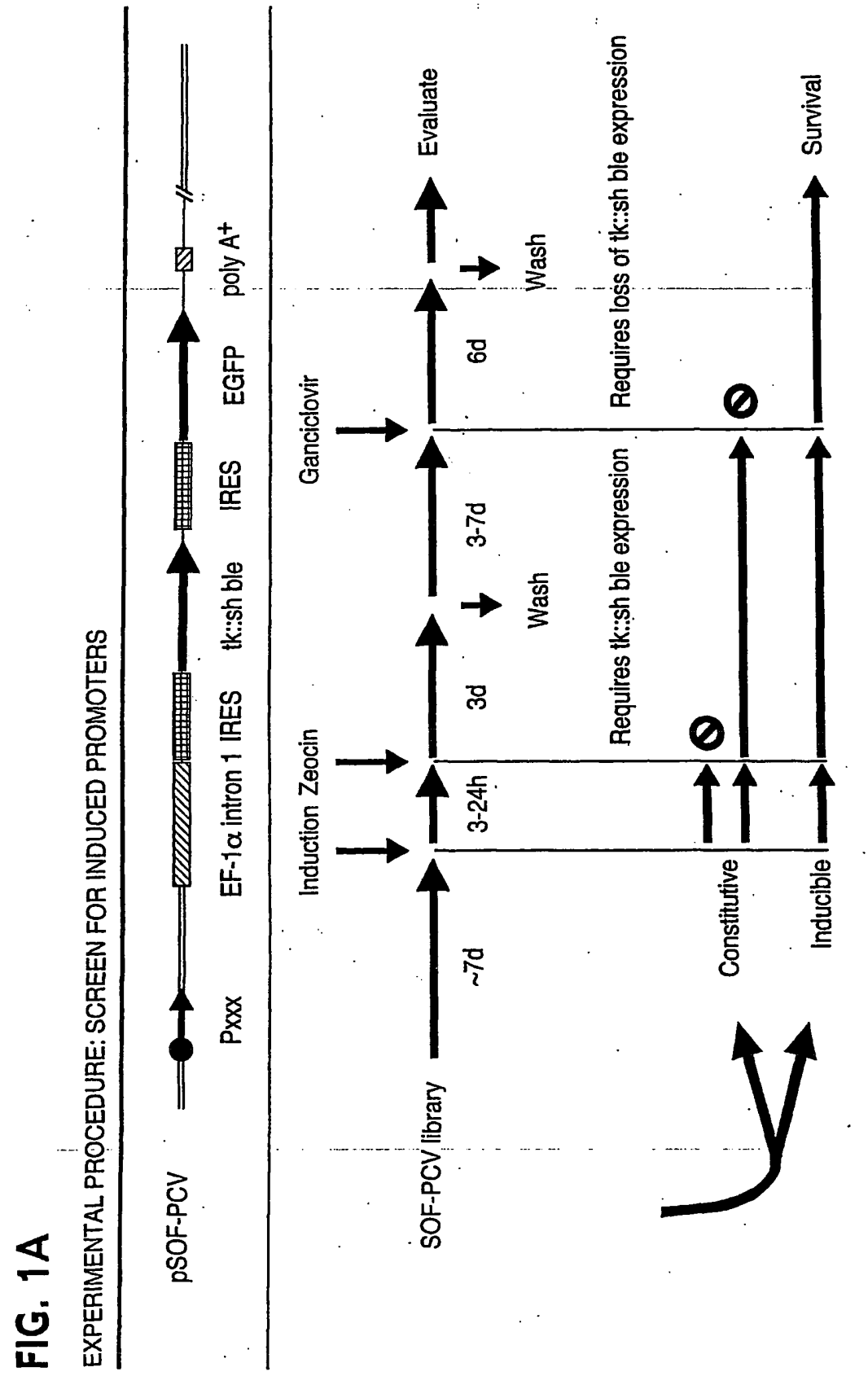

EXPERIMENTAL PROCEDURE: SCREEN FOR INDUCED PROMOTERS
pSOF-PCV
Pxxx
EF-1α intron 1
IRES
tk::sh ble
IRES
EGFP
poly A+
SOF-PCV library
~7d
Induction Zeocin
3-24h
3d
Wash
3-7d
Ganciclovir
6d
Wash
Evaluate
Requires tk::sh ble expression
Requires loss of tk::sh ble expression
Constitutive
Inducible
Survival

FIG. 3A

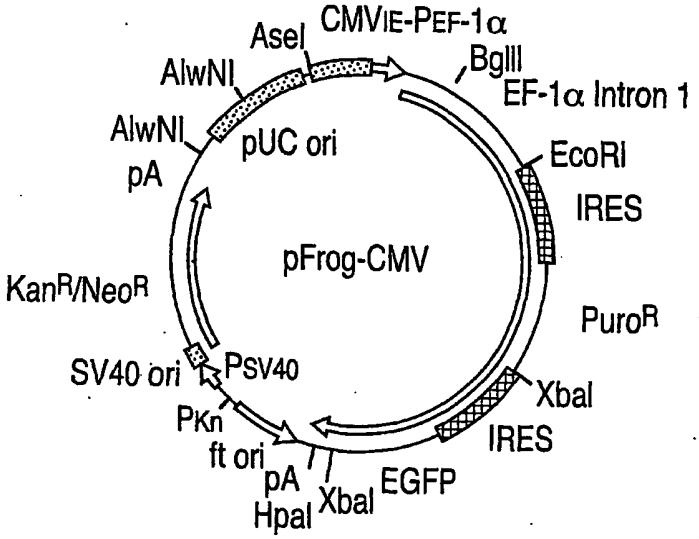
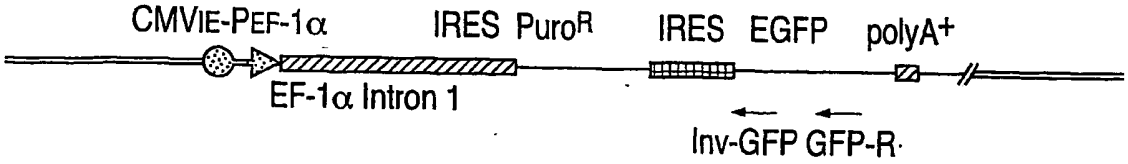

CMVIE-PEF-1α
AseI
BglII
AlwNI
EF-1α Intron 1
AlwNI
pUC ori
EcoRI
pA
IRES
pFrog-CMV
KanR/NeoR
PuroR
SV40 ori
PSV40
XbaI
PKn
IRES
ft ori
pA
XbaI
EGFP
HpaI

FIG. 3B


CMVIE-PEF-1α
IRES PuroR
IRES EGFP
polyA+
EF-1α Intron 1
Inv-GFP GFP-R

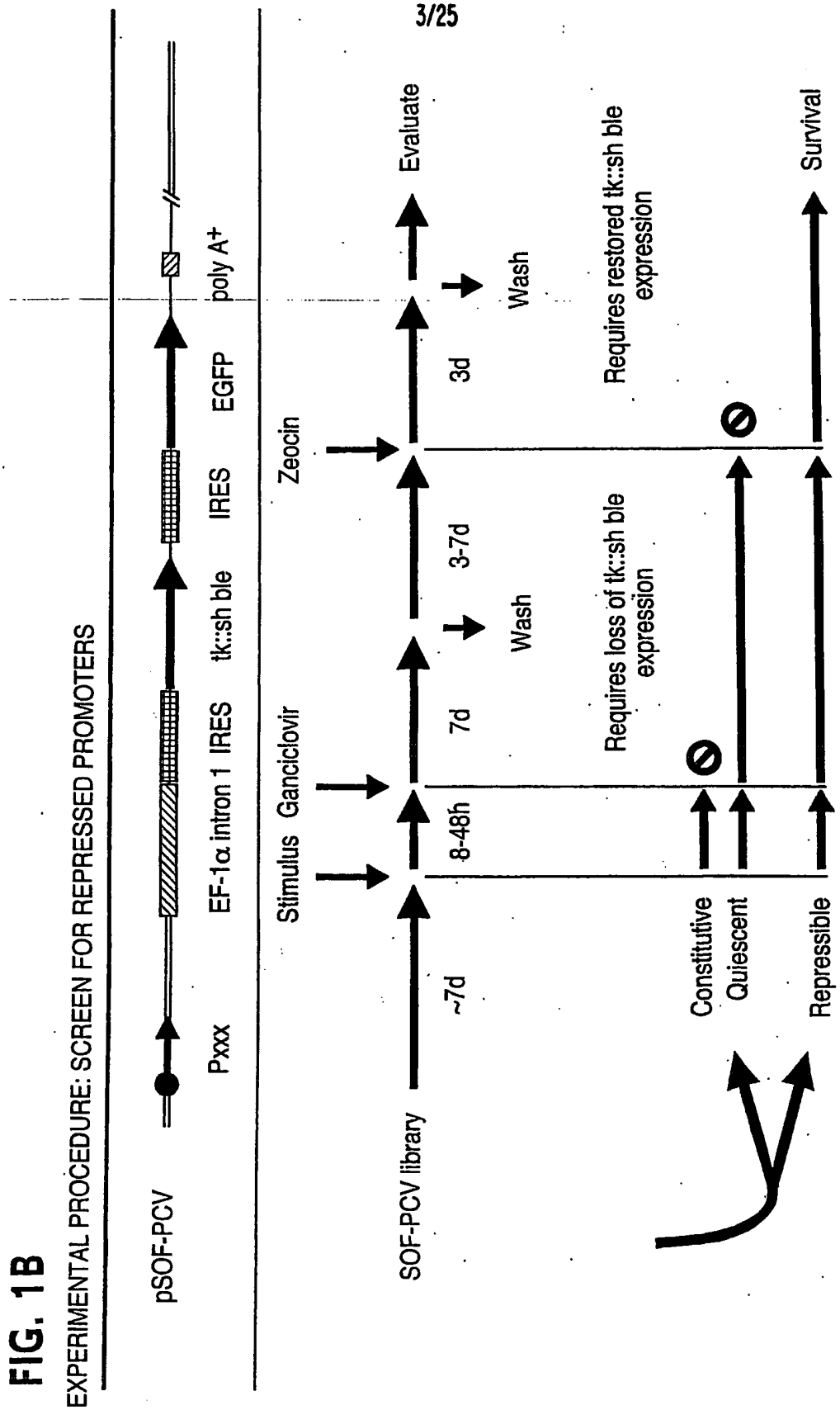


FIG. 1B

EXPERIMENTAL PROCEDURE: SCREEN FOR REPRESSED PROMOTERS

FIG. 2A

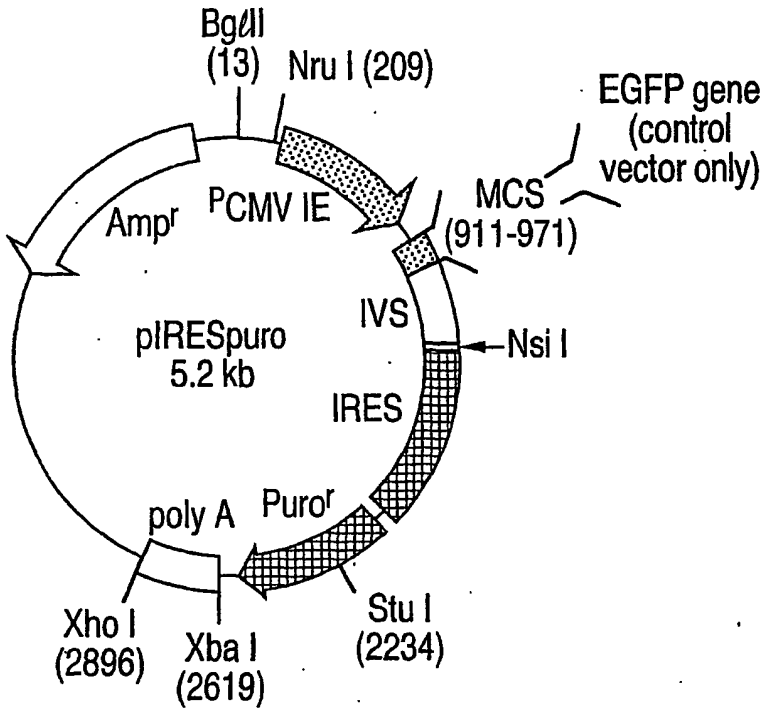
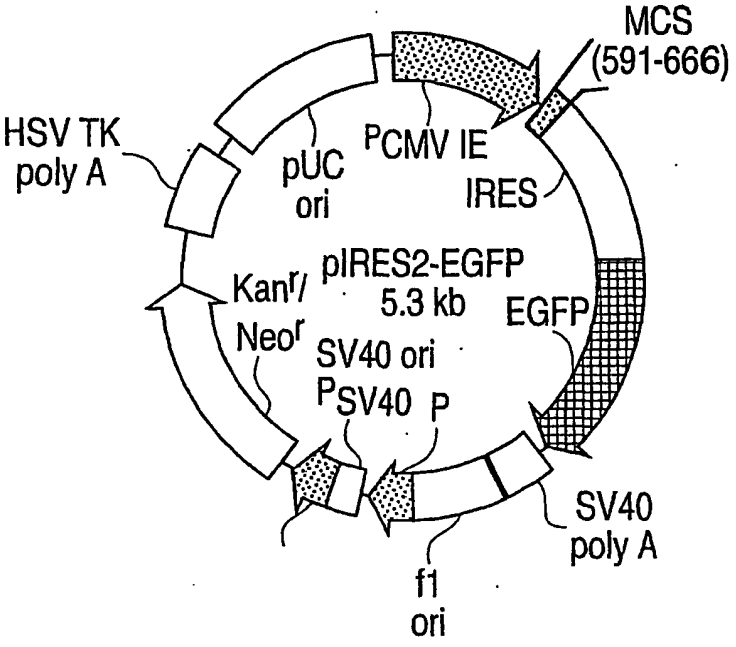

Bgl/II
(13)
Nru I (209)
EGFP gene
(control
vector only)
MCS
(911-971)
Ampr
PCMV IE
IVS
Nsi I
pIRESpuro
5.2 kb
IRES
poly A
Puror
Xho I
(2896)
Xba I
(2619)
Stu I
(2234)

FIG. 2B


MCS
(591-666)
HSV TK
poly A
pUC
ori
PCMV IE
IRES
pIRES2-EGFP
5.3 kb
Kanr/
Neor
EGFP
SV40 ori
PSV40
P
SV40
poly A
f1
ori

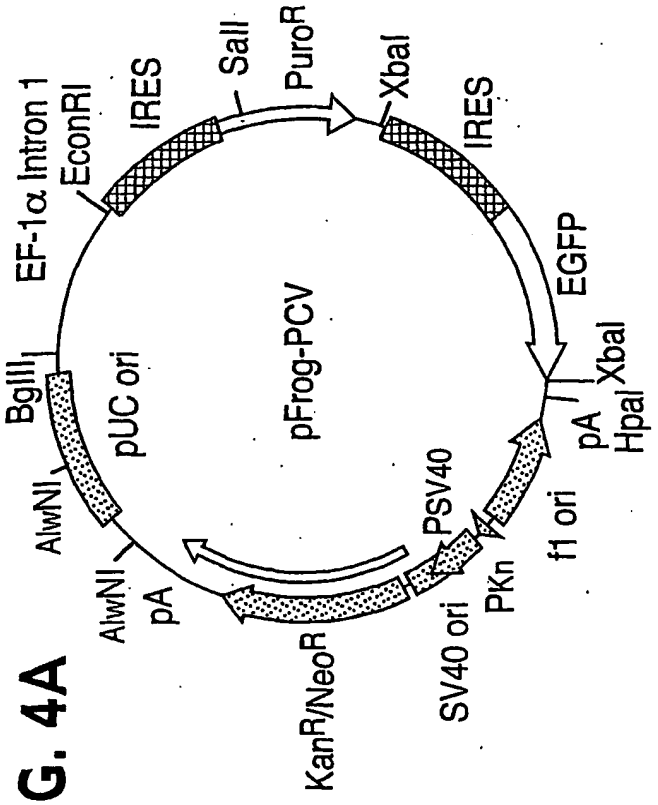

pFrog-PCV
EF-1α Intron 1
EconRI
IRES
SalI
PuroR
XbaI
IRES
EGFP
XbaI
HpaI
pA
f1 ori
PKn
PSV40
SV40 ori
KanR/NeoR
pA
AlwNI
AlwNI
pUC ori
BglII

FIG. 4A

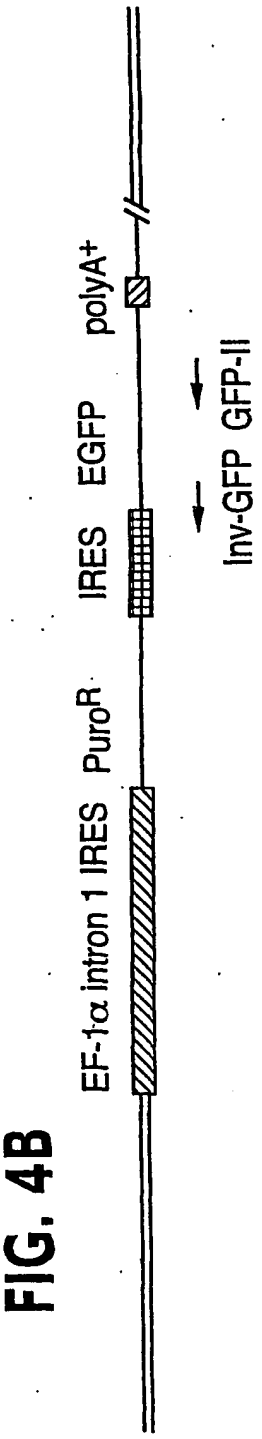

EF-1α intron 1 IRES PuroR
IRES EGFP polyA+
Inv-GFP GFP-II

FIG. 4B

FIG. 5

```
        10          20          30          40          50          60
atgcatacaa  ggagacgacc  ttccATGTCG  ACTACTAACC  TTCTTCTCtt  tcctacagCT
        70          80          90         100         110         120
TCGTACCCCG  GCCATCAACA  CGCGTCTGCG  TTCGACCAGG  CTGCGCGTTC  TCGCGGCCAT
       130         140         150         160         170         180
AGCAACCGAC  GTACGGCGTT  GCGCCCTCGC  CGGCAGCAAG  AAGCCACGGA  AGTCCGCCCG
       190         200         210         220         230         240
GAGCAGAAAA  TGCCCACGCT  ACTGCGGGTT  TATATAGACG  GTCCCCACGG  GATGGGGAAA
       250         260         270         280         290         300
ACCACCACCA  CGCAACTGCT  GGTGGCCCTG  GGTTCGCGCG  ACGATATCGT  CTACGTACCC
       310         320         330         340         350         360
GAGCCGATGA  CTTACTGGCG  GGTGCTGGGG  GCTTCCGAGA  CAATCGCGAA  CATCTACACC
       370         380         390         400         410         420
ACACAACACC  GCCTCGACCA  GGGTGAGATA  TCGGCCGGGG  ACGCGGCGGT  GGTAATGACA
       430         440         450         460         470         480
AGCGCCCAGA  TAACAATGGG  CATGCCTTAT  GCCGTGACCG  ACGCCGTTCT  GGCTCCTCAT
       490         500         510         520         530         540
ATCGGGGGGG  AGGCTGGGAG  CTCACATGCC  CCGCCCCCGG  CCCTCACCCT  CATCTTCGAC
       550         560         570         580         590         600
CGCCATCCCA  TCGCCGCCCT  CCTGTGCTAC  CCGGCCGCGC  GGTACCTTAT  GGGCAGCATG
       610         620         630         640         650         660
ACCCCCCAGG  CCGTGCTGGC  GTTCGTGGCC  CTCATCCCGC  CGACCTTGCC  CGGCACCAAC
       670         680         690         700         710         720
ATCGTGCTTG  GGGCCCTTCC  GGAGGACAGA  CACATCGACC  GCCTGGCCAA  ACGCCAGCGC
       730         740         750         760         770         780
CCCGGCGAGC  GGCTGGACCT  GGCTATGCTG  GCTGCGATTC  GCCGCGTTTA  CGGGCTACTT
       790         800         810         820         830         840
GCCAATACGG  TGCGGTATCT  GCAGTGCGGC  GGGTCGTGGC  GGGAGGACTG  GGGACAGCTT
       850         860         870         880         890         900
CCGGGGACGG  CCGTGCCGCC  CCAGGGTGCC  GAGCCCCAGA  GCAACGCGGG  CCCACGACCC
       910         920         930         940         950         960
TATATCGGGG  ACACGTTATT  TACCCTGTTT  CGGGCCCCCG  AGTTGCTGGC  CCCCAACGGC
       970         980         990        1000        1010        1020
AACCTGTATA  ACGTGTTTGC  CTGGGCCTTG  GACGTCTTGG  CCAAACGCCT  CCGTTCCATG

      1030        1040        1050        1060        1070        1080
CACGTCTTTA  TCCTGGATTA  CGACCAATCG  CCCGCCGGCT  GCCGGGACGC  CCTGCTGCAA
      1090        1100        1110        1120        1130        1140
CTTACCTCCG  GGATGGTCCA  GACCCACGTC  ACCACCCCCG  GCTCCATACC  GACGATATGC
      1150        1160        1170        1180        1190        1200
GACCTGGCGC  GCACGTTTGC  CCGTGAGATG  ATCAGCGGAG  CTAATGGCGT  CATGGCCAAG
      1210        1220        1230        1240        1250        1260
TTGACCAGTG  CCGTTCCGGT  GCTCACCGCG  CGCGACGTCG  CCGGAGCGGT  CGAGTTCTGG
      1270        1280        1290        1300        1310        1320
ACCGACCGGC  TCGGGTTCTC  CCGGGACTTC  GTGGAGGACG  ACTTCGCCGG  TGTGGTCCGG
      1330        1340        1350        1360        1370        1380
GACGACGTGA  CCCTGTTCAT  CAGCGCGGTC  CAGGACCAGG  TGGTGCCGGA  CAACACCCTG
      1390        1400        1410        1420        1430        1440
GCCTGGGTGT  GGGTGCGCGG  CCTGGACGAG  CTGTACGCCG  AGTGGTCGGA  GGTCGTGTCC
      1450        1460        1470        1480        1490        1500
ACGAACTTCC  GGGACGCCTC  CGGGCCGGCC  ATGACCGAGA  TCGGCGAGCA  GCCGTGGGGG
      1510        1520        1530        1540        1550        1560
CGGGAGTTCG  CCCTGCGCGA  CCCGGCCGGC  AACTGCGTGC  ACTTCGTGGC  CGAGGAGCAG
      1570        1580        1590        1600        1610        1620
GACTGAggat  cctctagatg  cat.....  ..........  ..........  ..........
```

FIG. 6A

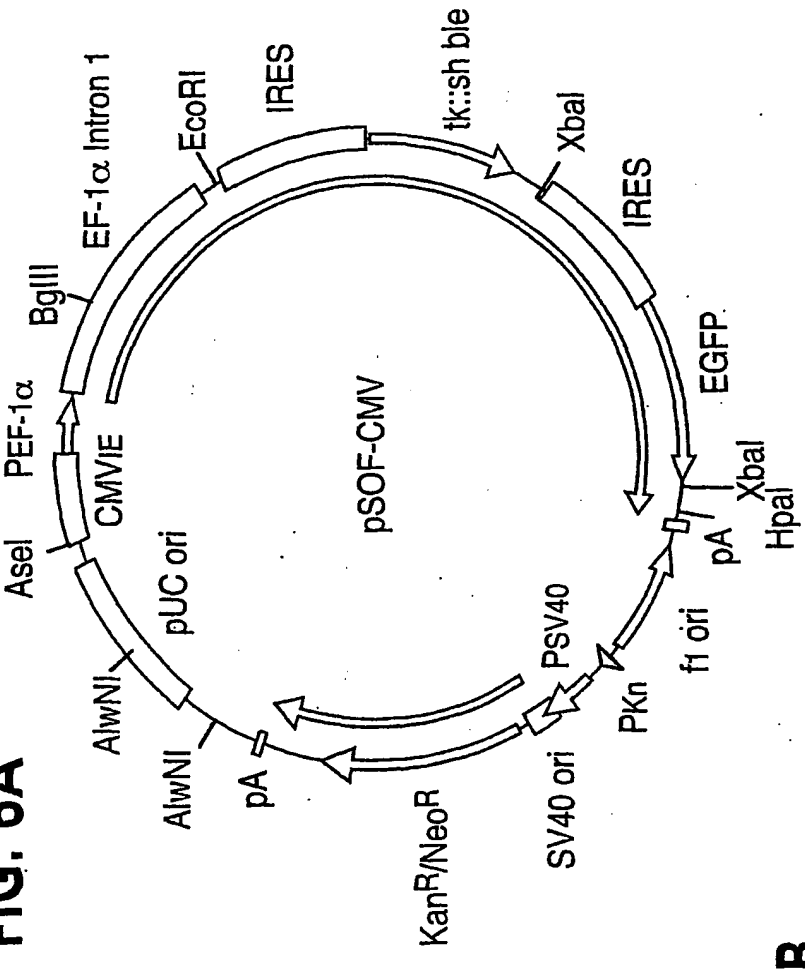

pSOF-CMV
AseI
PEF-1α
CMVIE
BglII
EF-1α Intron 1
EcoRI
IRES
tk::sh ble
XbaI
IRES
EGFP
XbaI
HpaI
pA
f1 ori
PKn
PSV40
SV40 ori
KanR/NeoR
pA
AlwNI
AlwNI
pUC ori

FIG. 6B

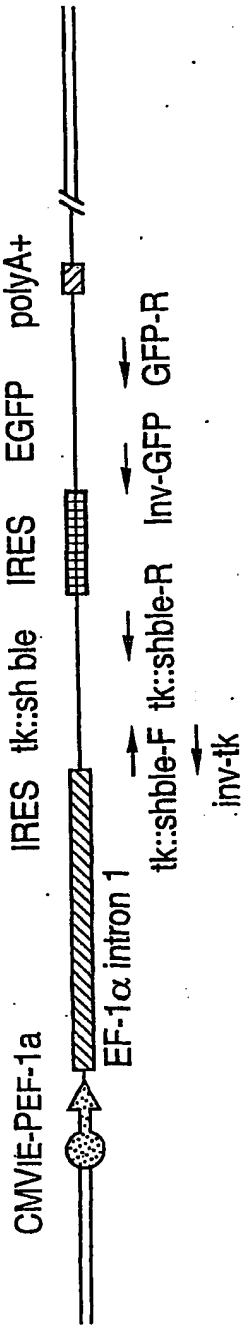

CMVIE-PEF-1a
EF-1α intron 1
IRES
tk::sh ble
IRES
EGFP
polyA+
tk::shble-F
tk::shble-R
Inv-GFP
GFP-R
inv-tk

FIG. 6A

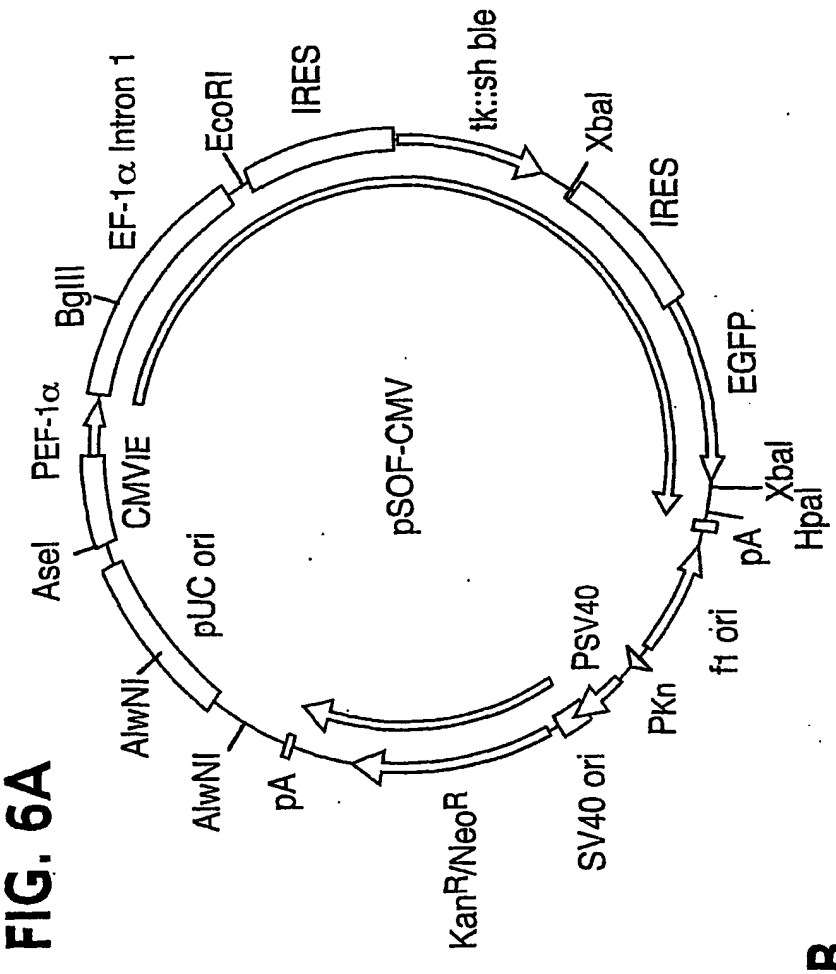
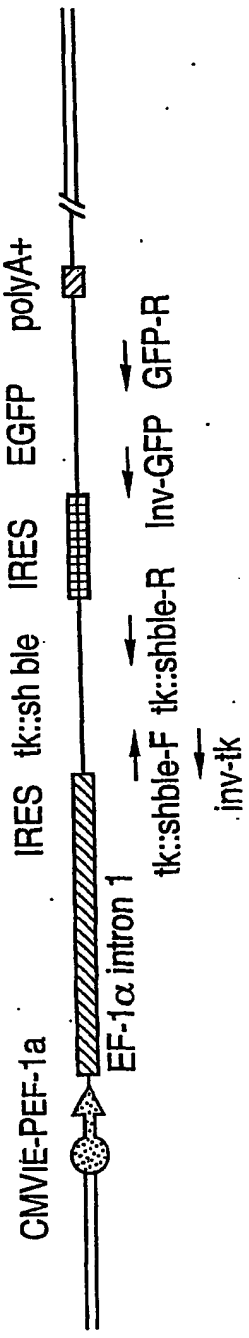

pSOF-CMV
AseI
PEF-1α
BglII
EF-1α Intron 1
EcoRI
IRES
tk::sh ble
XbaI
IRES
EGFP
XbaI
HpaI
pA
f1 ori
PKn
PSV40
SV40 ori
KanR/NeoR
pA
AlwNI
AlwNI
pUC ori
CMVIE

FIG. 6B


CMVIE-PEF-1a
EF-1α intron 1
IRES
tk::sh ble
tk::shble-F
tk::shble-R
inv-tk
IRES
EGFP
Inv-GFP
GFP-R
polyA+

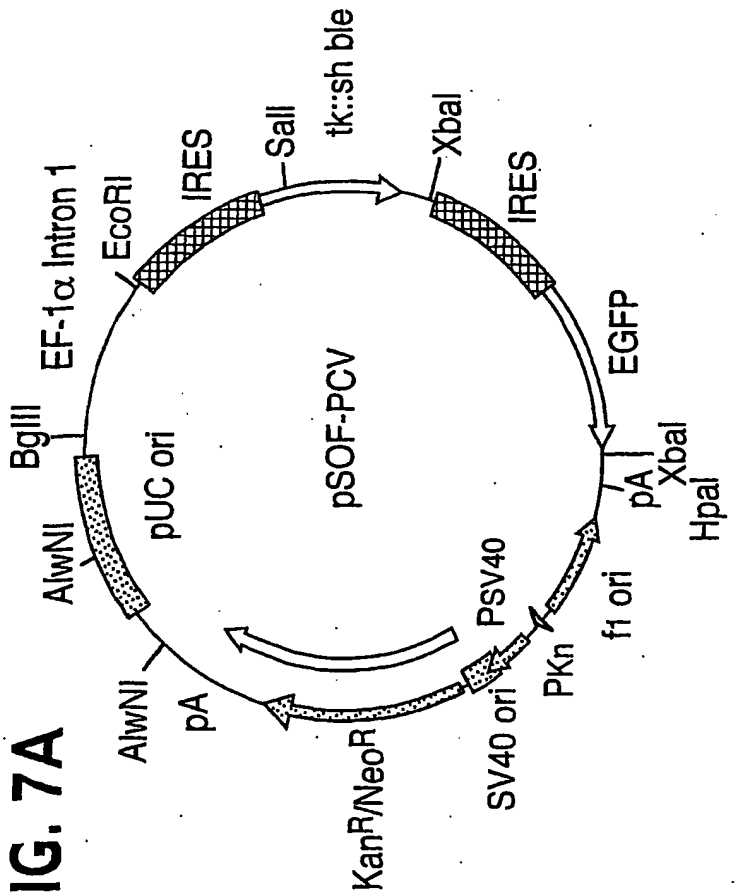

pSOF-PCV
BglII
EF-1α Intron 1
EcoRI
IRES
SalI
tk::sh ble
XbaI
IRES
EGFP
XbaI
HpaI
pA
f1 ori
PKn
PSV40
SV40 ori
KanR/NeoR
pA
AlwNI
AlwNI
pUC ori

FIG. 7A

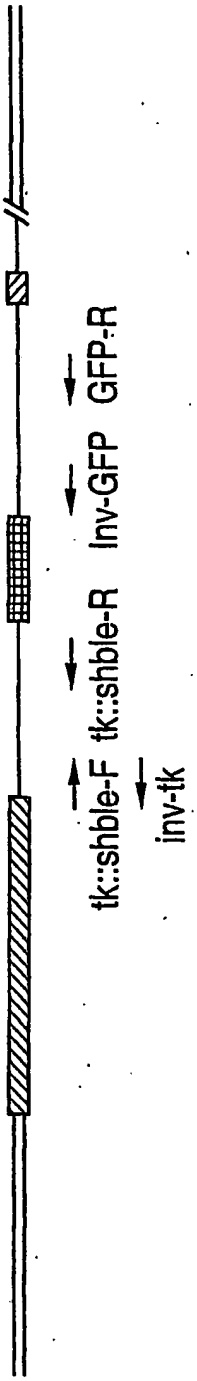

EF-1α intron 1
IRES
tk::sh ble
IRES
EGFP
polyA+
tk::shble-F
tk::shble-R
Inv-GFP
GFP-R
inv-tk

FIG. 7B

FIG. 8A

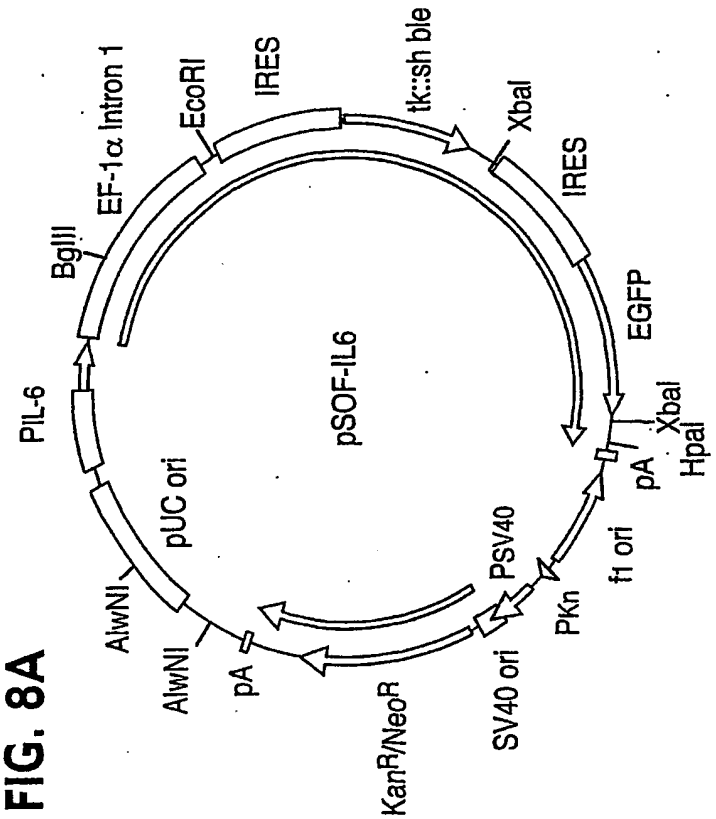

PIL-6
BglII
EF-1α Intron 1
EcoRI
IRES
tk::sh ble
XbaI
IRES
EGFP
XbaI
HpaI
pA
f1 ori
PKn
PSV40
SV40 ori
KanR/NeoR
pA
AlwNI
AlwNI
pUC ori
pSOF-IL6

FIG. 8B

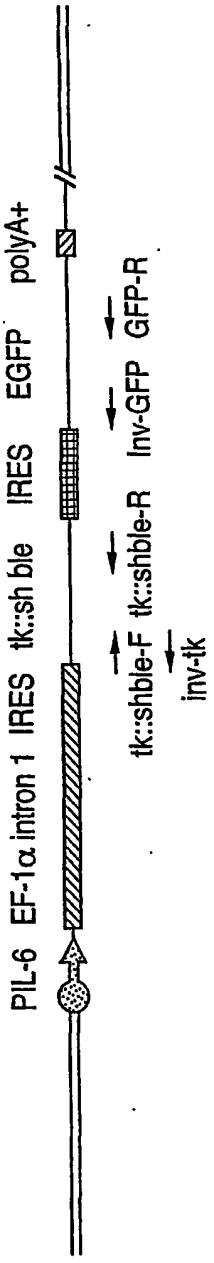
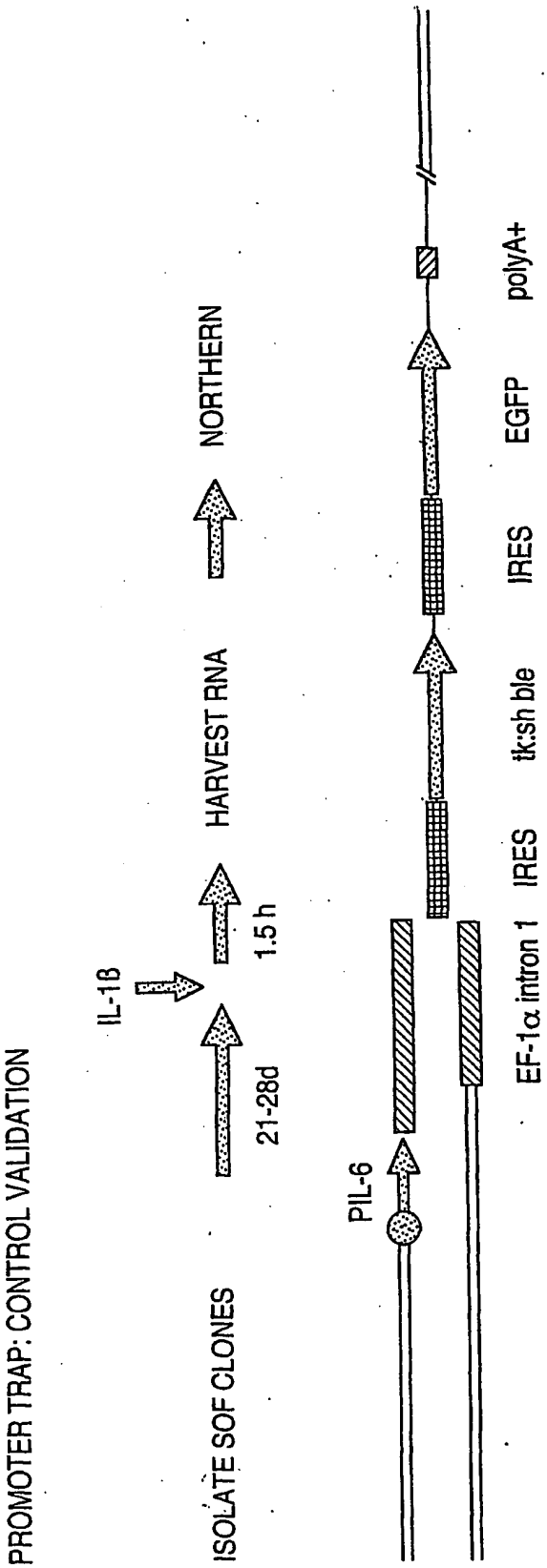

PIL-6
EF-1α intron 1
IRES
tk::sh ble
IRES
EGFP
polyA+
tk::shble-F
tk::shble-R
inv-tk
Inv-GFP
GFP-R

FIG. 9A


PROMOTER TRAP: CONTROL VALIDATION
ISOLATE SOF CLONES
21-28d
IL-1β
1.5 h
HARVEST RNA
NORTHERN
PIL-6
EF-1α intron 1
IRES
tk:sh ble
IRES
EGFP
polyA+

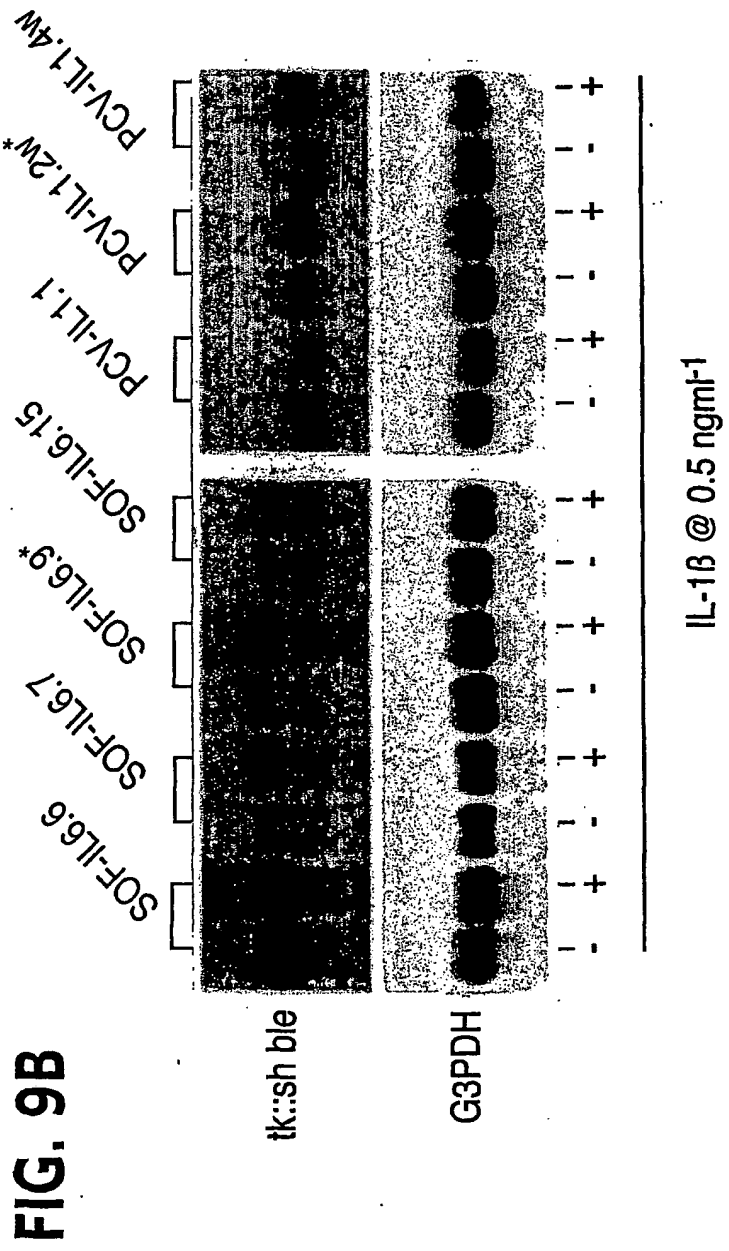

pCV-IL1.4w
pCV-IL1.2w*
pCV-IL1.1
SOF-IL6.15
SOF-IL6.9*
SOF-IL6.7
SOF-IL6.6
tk::sh ble
G3PDH
IL-1β @ 0.5 ngml-1
+
-

FIG. 9B

FIG. 9C

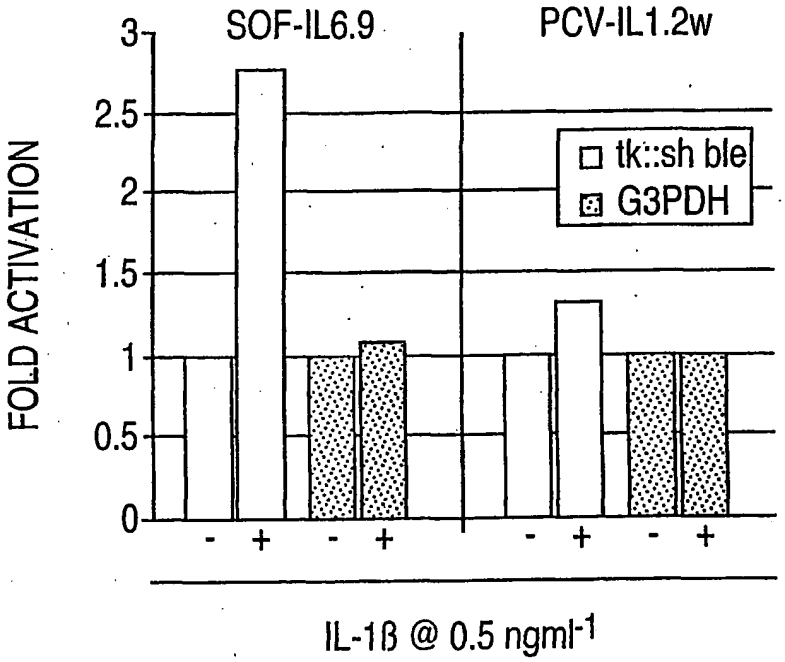

SOF-IL6.9
PCV-IL1.2w
FOLD ACTIVATION
3
2.5
2
1.5
1
0.5
0
tk::sh ble
G3PDH
- + - +
- + - +
IL-1ß @ 0.5 ngml-1

FIG. 10A

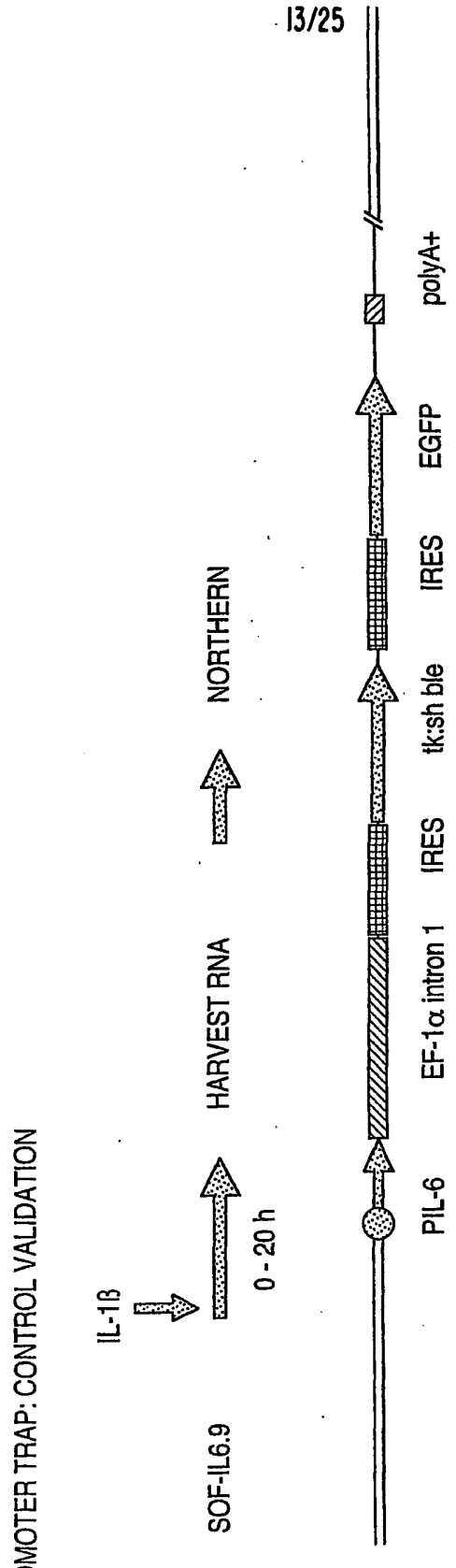

PROMOTER TRAP: CONTROL VALIDATION
SOF-IL6.9
IL-1ß
0 - 20 h
HARVEST RNA
NORTHERN
PIL-6
EF-1α intron 1
IRES
tk:sh ble
IRES
EGFP
polyA+

FIG. 10B

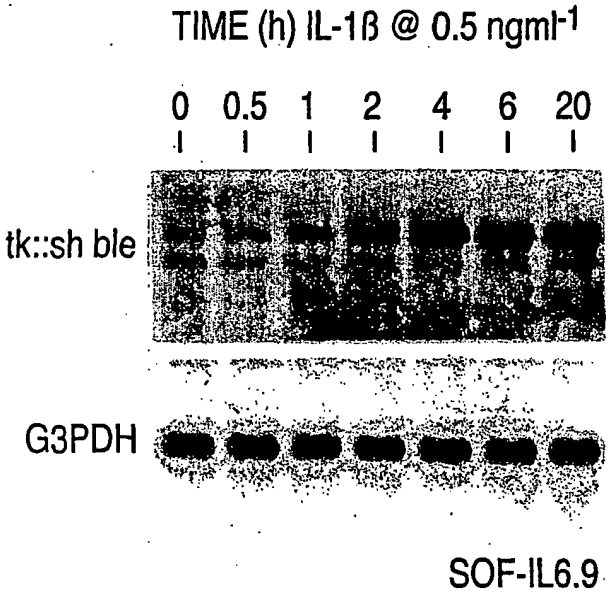


FIG. 10C

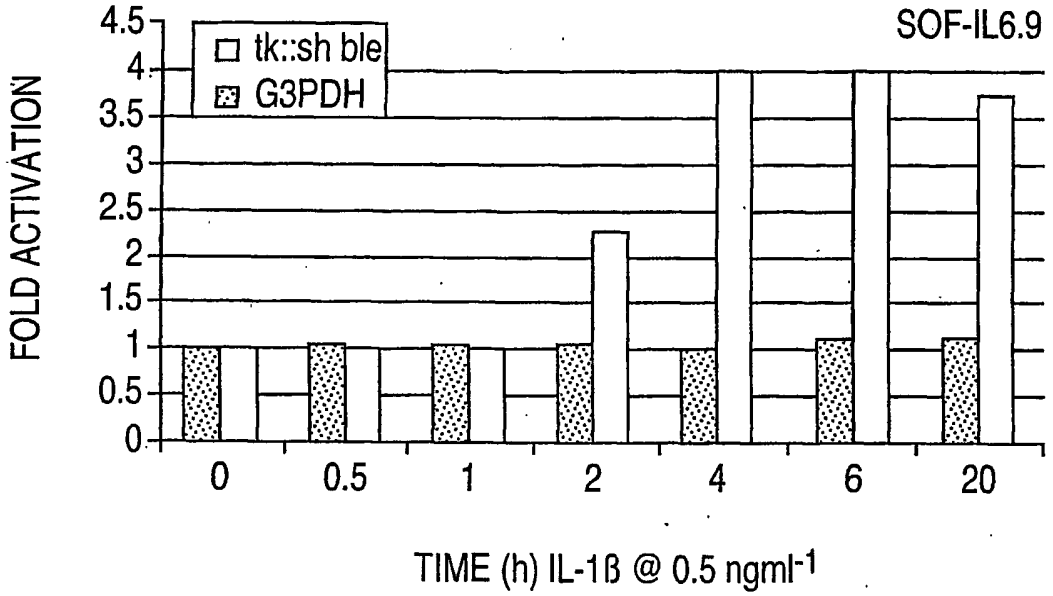

SOF-IL6.9
tk::sh ble
G3PDH
FOLD ACTIVATION
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
2
4
6
20
TIME (h) IL-1ß @ 0.5 ngml-1

FIG. 11A

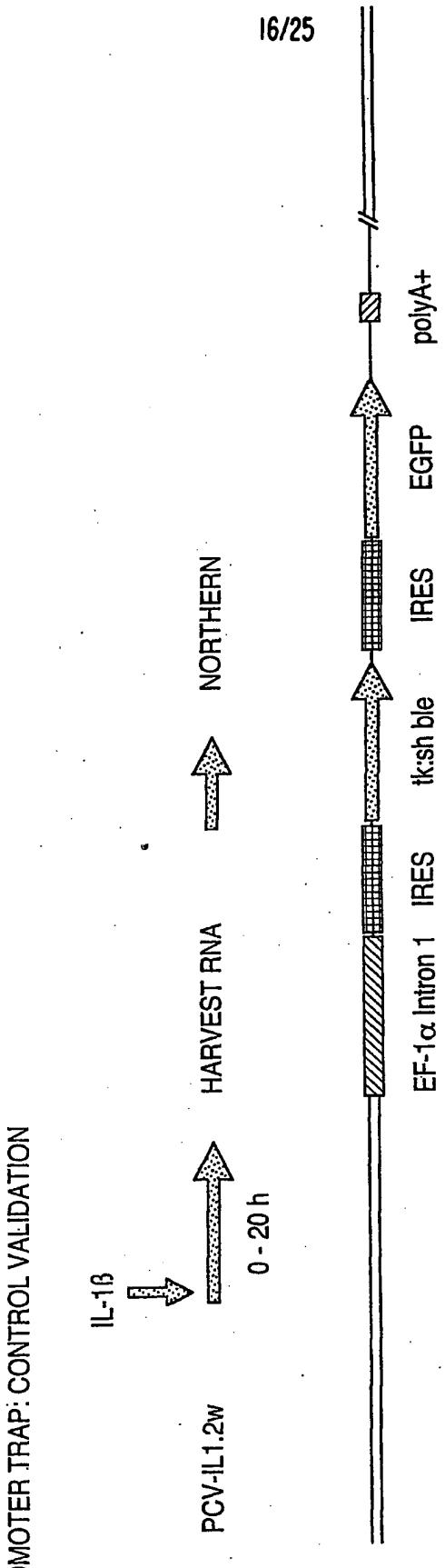

PROMOTER TRAP: CONTROL VALIDATION
IL-1β
PCV-IL1.2w
0 - 20 h
HARVEST RNA
NORTHERN
EF-1α Intron 1
IRES
tk:sh ble
IRES
EGFP
polyA+

FIG. 11B

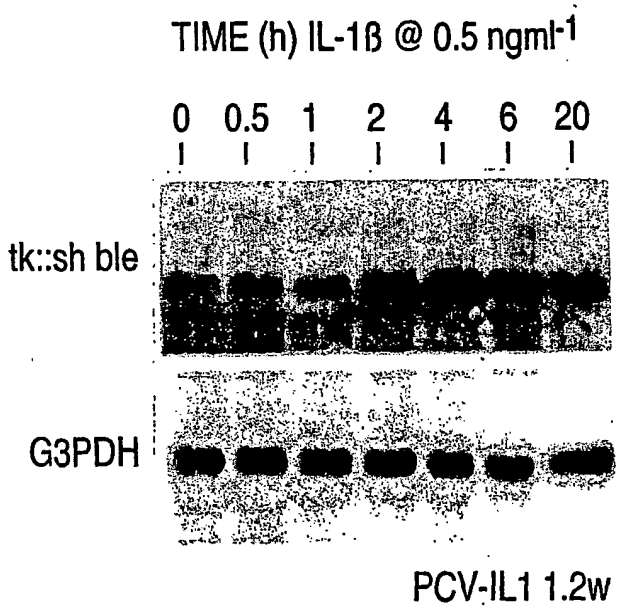
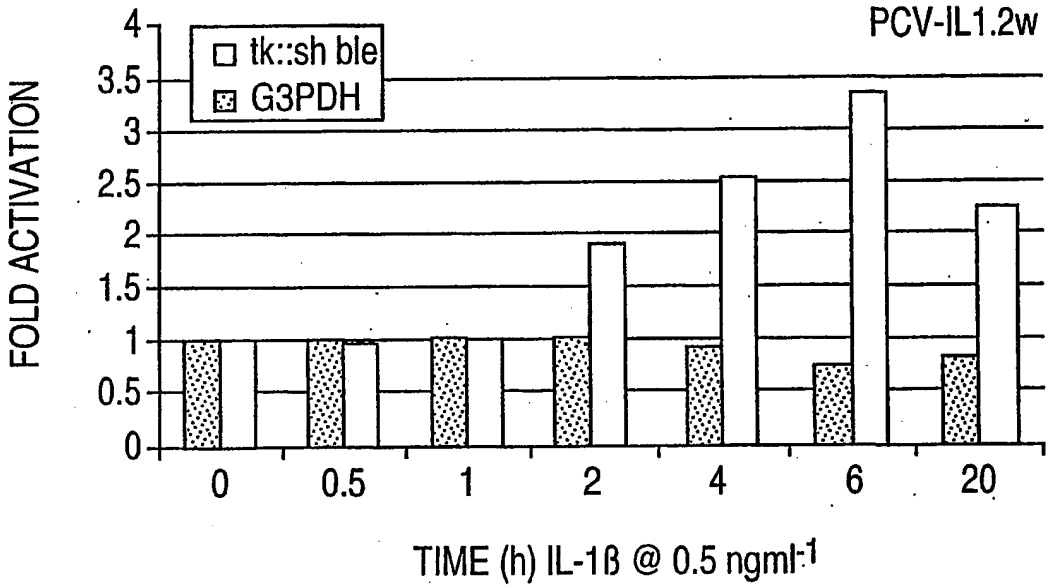


FIG. 11C


PCV-IL1.2w
tk::sh ble
G3PDH
FOLD ACTIVATION
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.5
1
2
4
6
20
TIME (h) IL-1ß @ 0.5 ngml-1

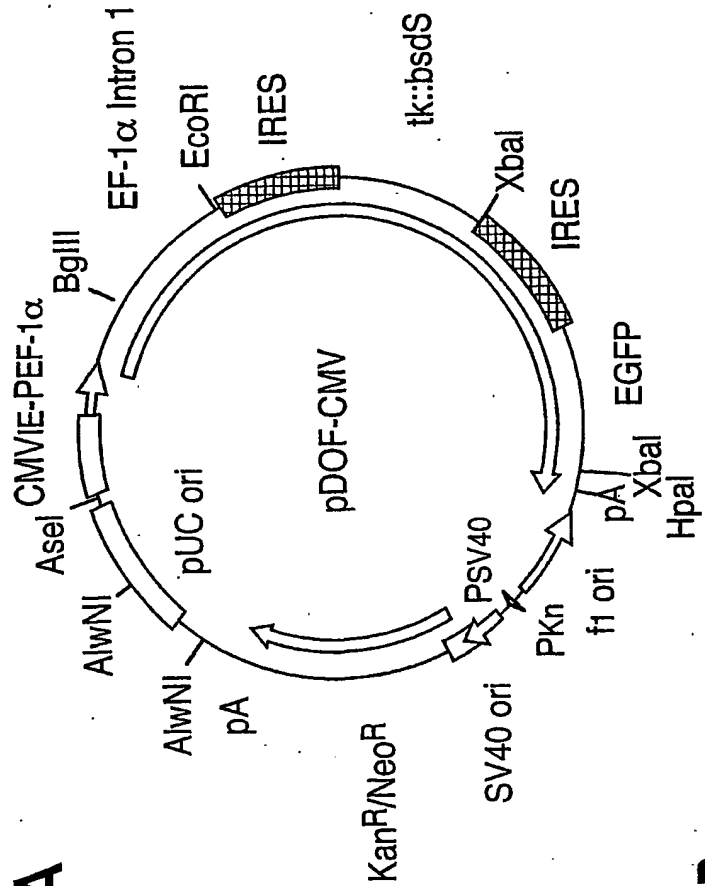


FIG. 12A

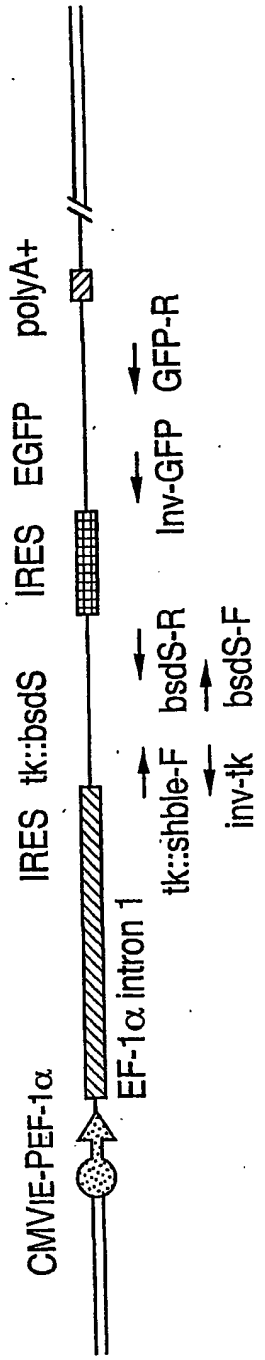


FIG. 12B

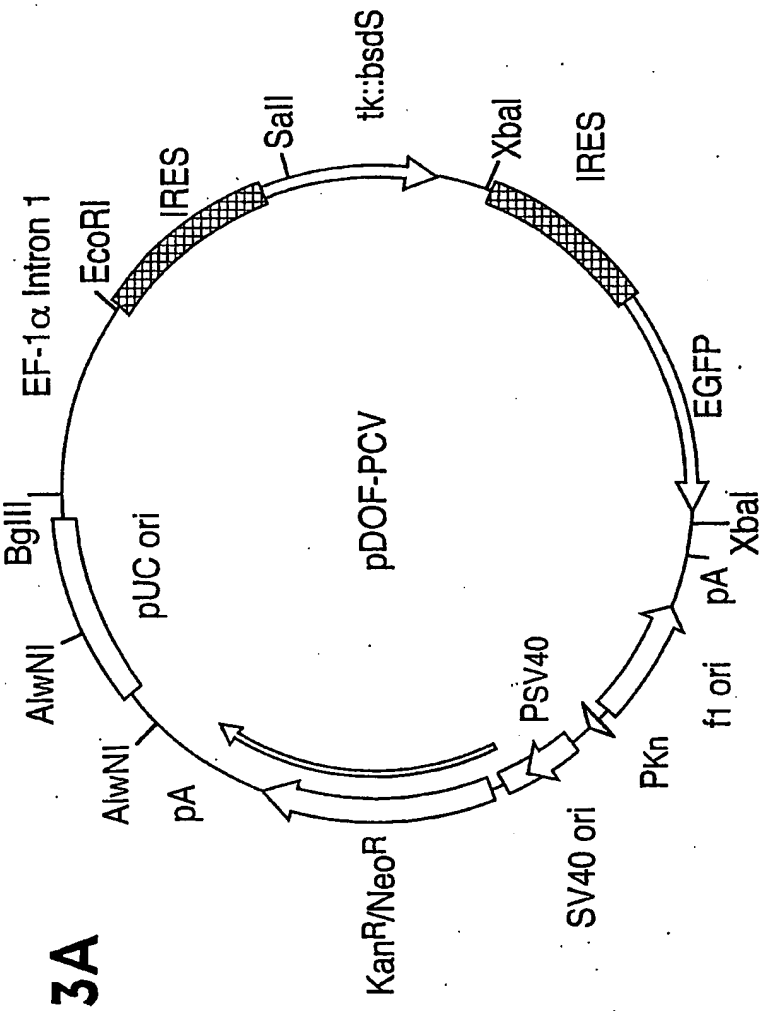


FIG. 13A

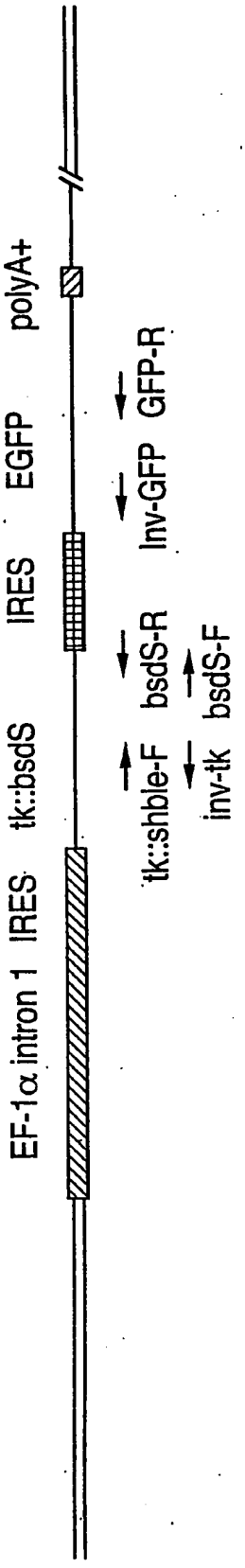


FIG. 13B

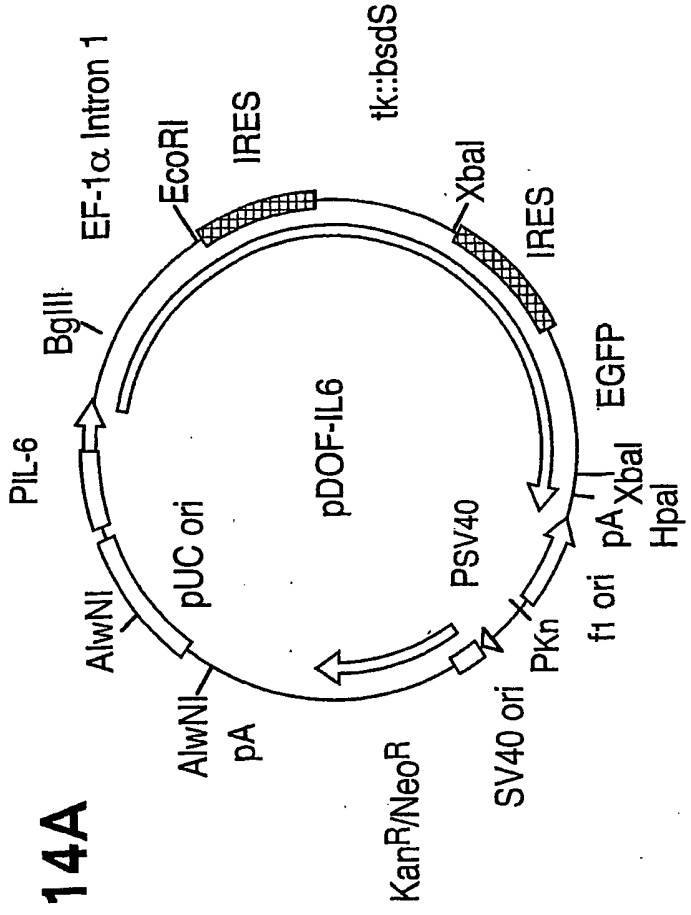


FIG. 14A

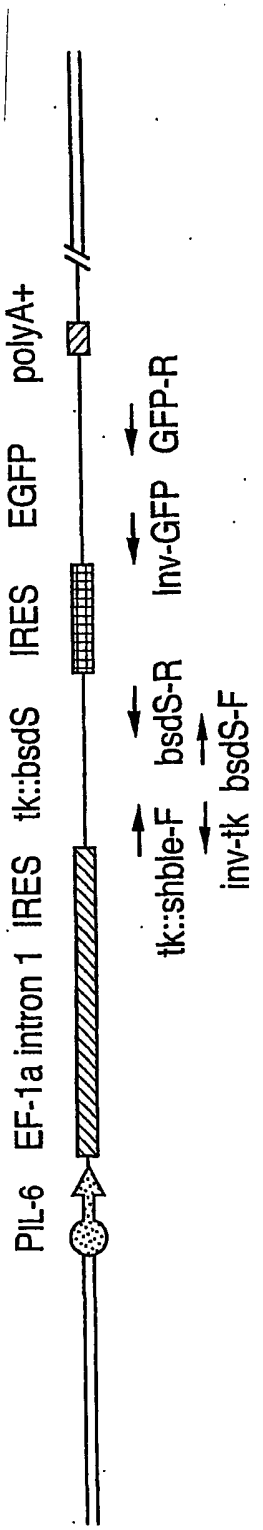


FIG. 14B

FIG. 15A

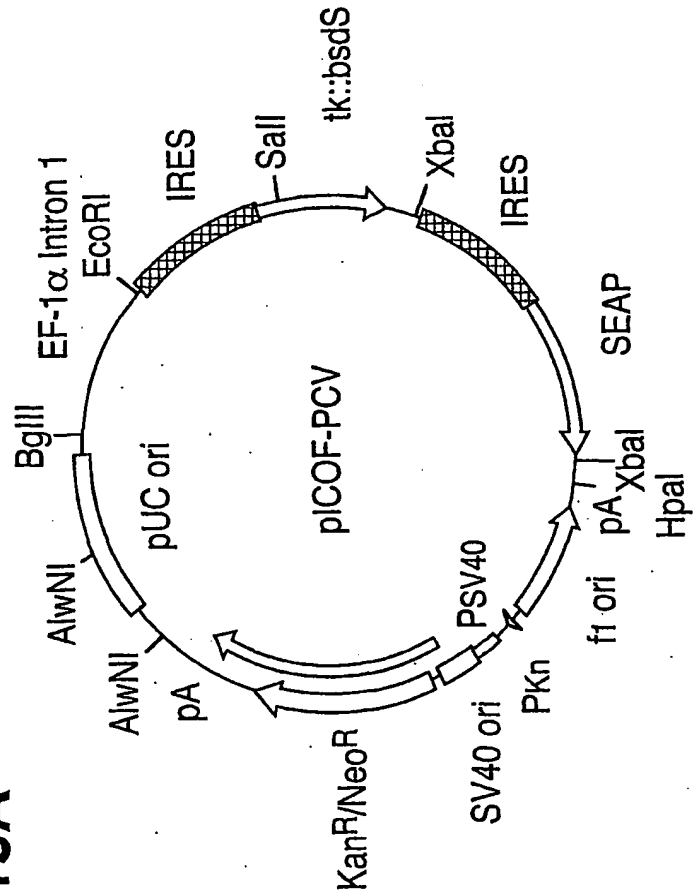
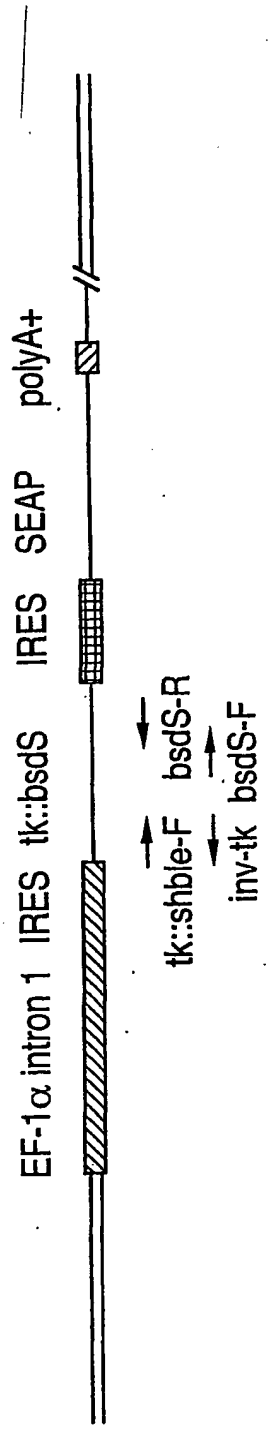


FIG. 15B



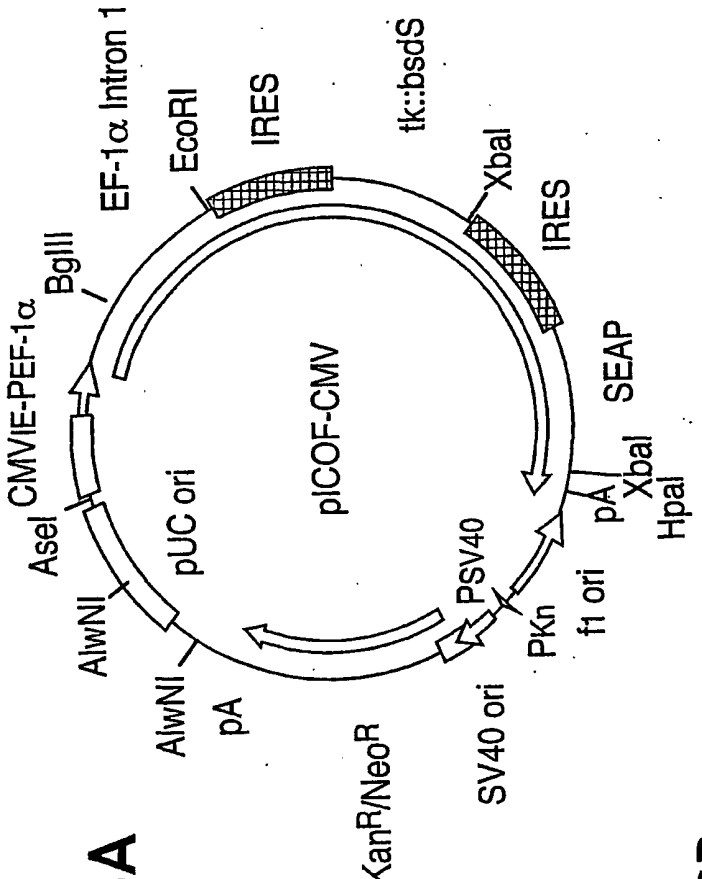


FIG. 16A

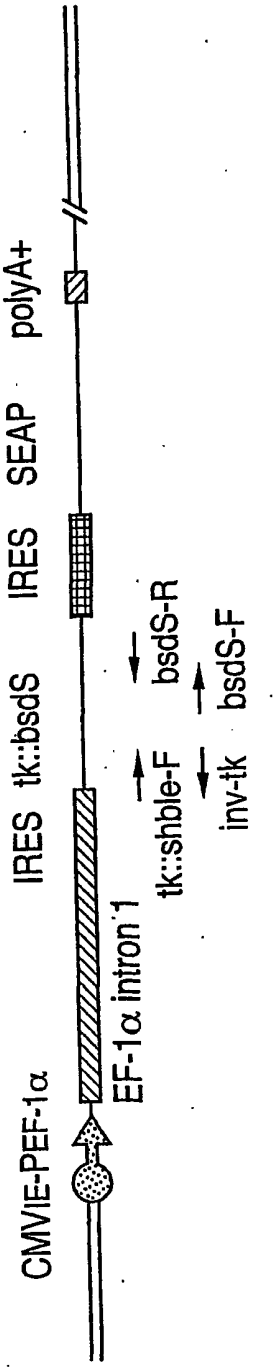


FIG. 16B

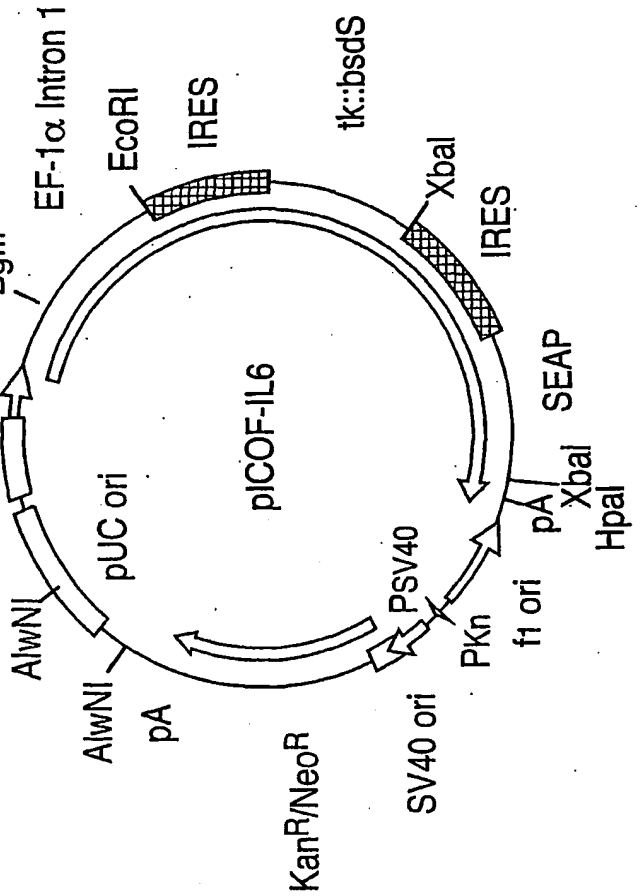


FIG. 17A

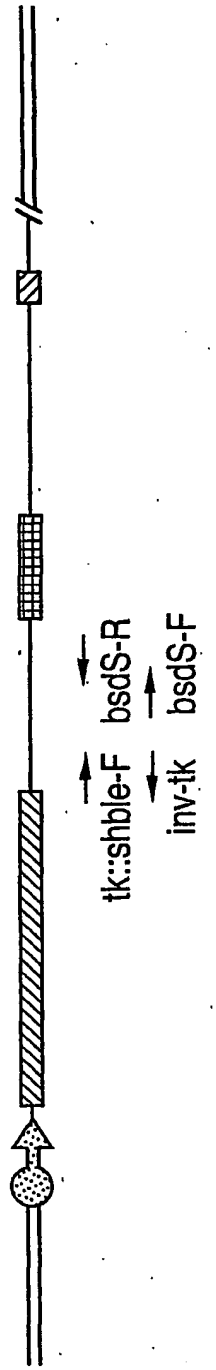
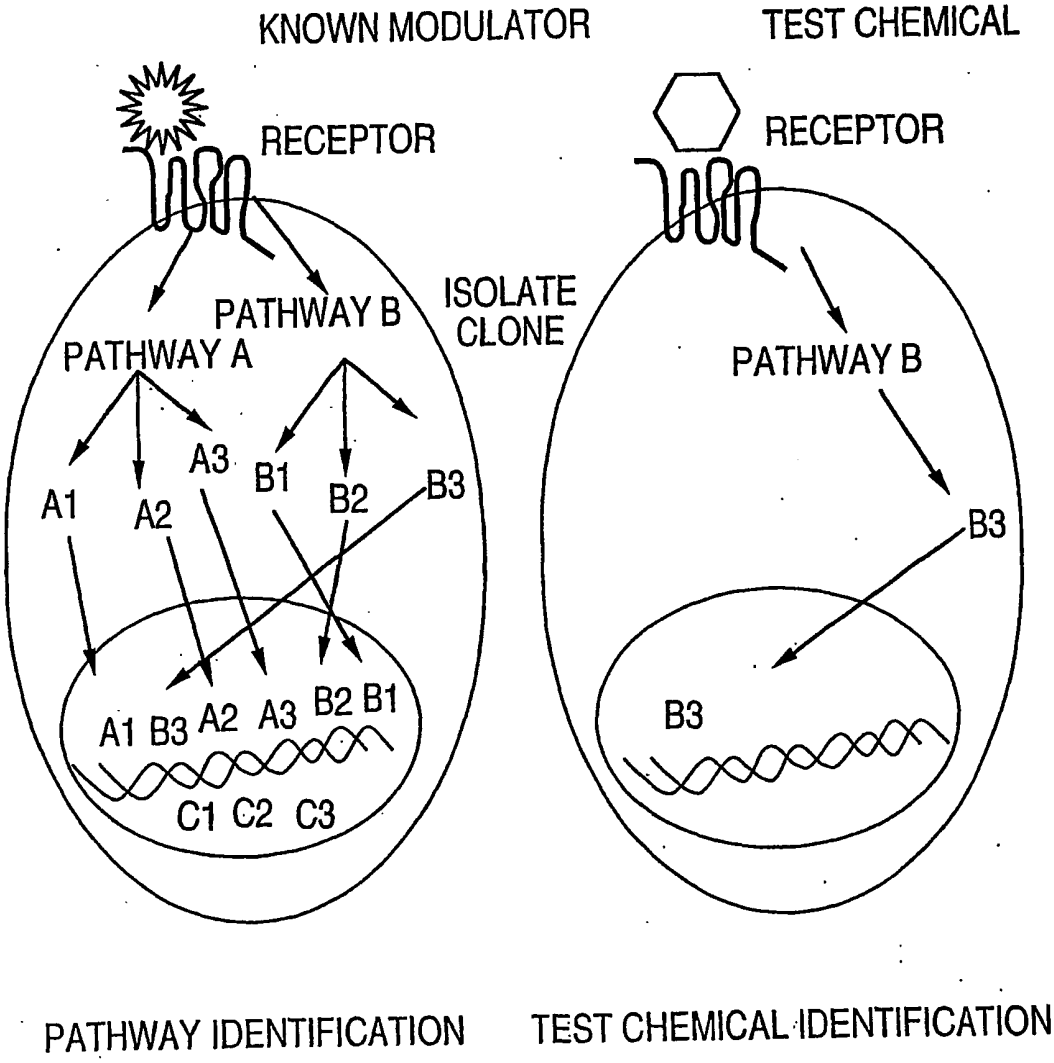


FIG. 17B

FIG. 18


GENOMIC ASSAY SYSTEM (GAS)
KNOWN MODULATOR
TEST CHEMICAL
RECEPTOR
RECEPTOR
PATHWAY B
PATHWAY A
ISOLATE CLONE
PATHWAY B
A1
A2
A3
B1
B2
B3
B3
A1 B3 A2 A3 B2 B1
C1 C2 C3
B3
PATHWAY IDENTIFICATION
TEST CHEMICAL IDENTIFICATION

INTERNATIONAL SEARCH REPORT

Interna l Application No
PCT/US 01/01480

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12N15/10 C12N15/62 C12N15/63 C12N15/64 C12N15/65
C12N5/10 C12Q1/68 G01N33/50

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 C12N C12Q G01N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
WPI Data, PAJ, CAB Data, STRAND, BIOSIS, EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 0 902 092 A (GSF FORSCHUNGSZENTRUM UMWELT ;CENTRE NAT RECH SCIENT (FR)) 17 March 1999 (1999-03-17) page 6, line 20 - line 39; claims 1-17; figures 1,8 page 7, line 1 - line 31 ---	29-32
Y	WO 98 13353 A (AURORA BIOSCIENCES CORP ;WHITNEY MICHAEL A (US); CRAIG FRANK (US);) 2 April 1998 (1998-04-02) cited in the application BLEC-2 claims 1-131; figures 1,3; example 1 --- -/--	1-9, 29-32

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
31 May 2001	19/06/2001

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Hornig, H

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
Y	HARRISON R W ET AL: "FUNCTIONAL IDENTIFICATION OF GENES UP- AND DOWN-REGULATED BY GLUCOCORTICOIDS IN ATT-20 PITUITARY CELLS USING AN ENHANCER TRAP" ENDOCRINOLOGY,US,BALTIMORE, MD, vol. 137, no. 7, 1 July 1996 (1996-07-01), pages 2758-2765, XP002068233 ISSN: 0013-7227 the whole document	1-9, 29-32
A	BONALDO PAOLO ET AL: "Efficient gene trap screening for novel developmental genes using IRESbetageo vector and in vitro preselection." EXPERIMENTAL CELL RESEARCH, vol. 244, no. 1, 10 October 1998 (1998-10-10), pages 125-136, XP002168648 ISSN: 0014-4827 cited in the application the whole document	
A	GOGOS JOSEPH A ET AL: "Selection for retroviral insertions into regulated genes." JOURNAL OF VIROLOGY, vol. 71, no. 2, 1997, pages 1644-1650, XP002168649 ISSN: 0022-538X cited in the application the whole document	
A	KARREMAN C: "A new set of positive/negative selectable markers for mammalian cells" GENE,NL,ELSEVIER BIOMEDICAL PRESS. AMSTERDAM, vol. 218, no. 1-2, September 1998 (1998-09), pages 57-61, XP004149340 ISSN: 0378-1119 cited in the application the whole document	
P,A	KUIPER MARCEL ET AL: "Cloning and characterization of a retroviral plasmid, pCC1, for combination suicide gene therapy." BIOTECHNIQUES, vol. 28, no. 3, March 2000 (2000-03), pages 572-576, XP002168650 ISSN: 0736-6205 the whole document	

INTERNATIONAL SEARCH REPORT

nation on patent family members

Interna J Application No
PCT/US 01/01480

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0902092 A	17-03-1999	DE 19740578 A	01-04-1999
		JP 11187876 A	13-07-1999
WO 9813353 A	02-04-1998	US 5928888 A	27-07-1999
		AU 4505797 A	17-04-1998
		EP 0952976 A	03-11-1999
		US 6107477 A	22-08-2000

THIS PAGE BLANK (USPTO)